AIWolfDial 2023: Summary of Natural Language Division of 5th International AIWolf Contest

INLG 2023 genChal @ Prague

Yoshinobu Kano

Claus Aranha, Kei Harada, Michimasa Inaba, Daisuke Katagami, Hirotaka Osawa, Takashi Otsuki, Fujio Toriumi

(AIWolf project organizers in alphabetical order)

Kano Laboratory

National University Corporation
Shizuoka University

One calm, quiet night

- This is a story about a small, sleepy village.
- Werewolves have arrived, who can change into humans and eat them.
- During the day, the werewolves have the same form as humans, at night they attack the villagers, one by one.
- Fear, uncertainty, and doubt towards the werewolves begin to grow.
- The villagers decide that they must execute those who are suspected of being werewolves, one by one...

The cover story of Are You a Werewolf?

Are You a Werewolf?

- “Are you a Werewolf?” is a party game that models a conflict between an informed minority and an uninformed majority
  - Initially, each player is secretly assigned a role affiliated with one of these two teams: Werewolf and Villager
- There are two phases: night and day
  - At night, the werewolves “attack” the villagers
  - During the day, surviving players (of both teams) discuss the elimination of a werewolf-suspect by voting
  - i.e. talk -> vote -> attack -> … until either team remains
- The teams have different objectives:
  - Villagers (Humans): To ascertain who the werewolves are, and kill them by votes
  - Werewolves: To kill off all the villagers without being killed themselves

What is “AIWolf”?

- A project to create AI agents that can play Werewolf (aka Mafia)
- Ability of Werewolf agents:
  - Negotiate, Convince, Plan, Lie...
- Key characteristics:
  - Social game
  - Imperfect information Game
  - Cooperative, Multi player
  - Rules Light / Ability to manipulate social capital
- One of the most challenging task in text generation/dialog system
  - Also a new frontier on Game AI and HCI

Challenge of AIWolf Contests

- Need to estimate knowledge and intentions of other agents
- And everything through dialogue only
  - Human-played games can be face-to-face
- As an evaluation of dialog systems
  - AI werewolf requires to deeply understand conversation to be a game
  - By evaluating links between dialog-game rather than win/lose
  - Any conversation available in games but should be game oriented, thus both task-oriented and non-task-oriented
- We can reveal current achievements and limitations of the generative AIs!
- We have been holding our annual open shared task contests for years...

C ests

- N
  in
- A
  -
- W
  lin
- W
  sh

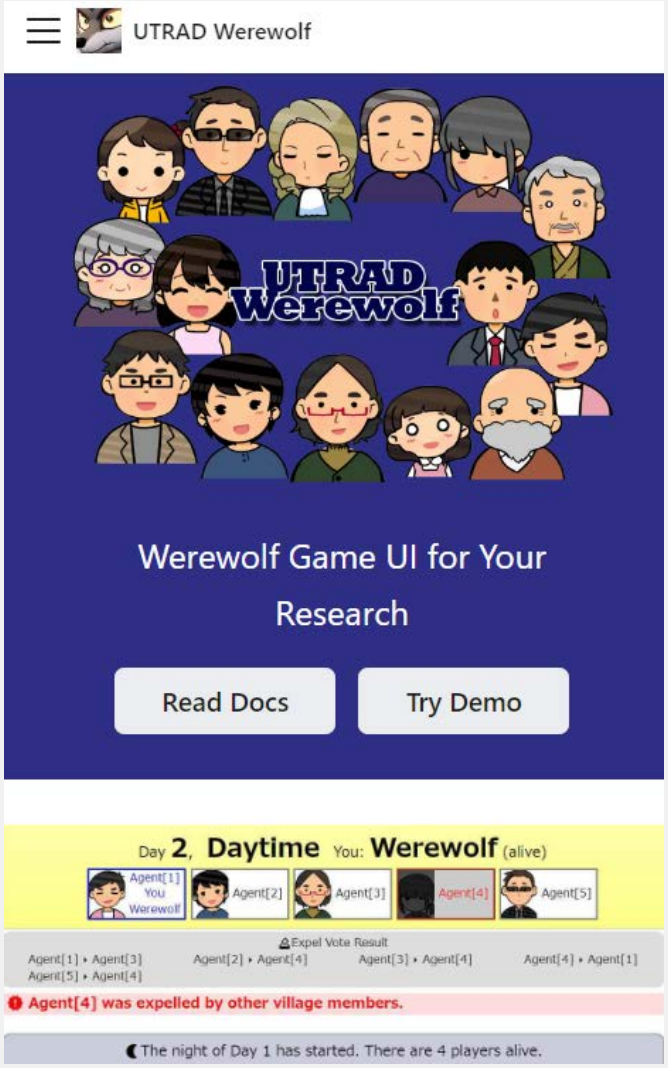
Three AIWolf Tasks (Divisions)

- Infrastructure Task
  - Submission of useful tools for the competition
- Protocol Task
  - Agents communicate using a fixed protocol; Evaluation by winning rate
- Natural Language Task (main report)
  - Agents communicate freely; Evaluation of communication quality by Judge panel
- Details in the following slides...

Infrastructure Division

This year's entries

- aiwolf-python by AIWolfSharp
  Python package for creating AIWolf agent.
- AIWolf Analyze by aiwolf team-ioh
  Web app for analyzing AIWolf game logs.
- UTRAD Werewolf by smikitky
  Web UI for the Werewolf game.
  - The Best Contribution Award
    voted by participants and organizers



Protocol Division

- Agents communicate via the AIWolf Protocol manually designed by the organizers
  - e.g. COMINGOUT Agent[03] SEER
- Evaluated on winning rate in 5- and 15-player games
- Top 15 from 33 teams selected in qualifying round
  - 2023-07-22/08-04, 244,000 games/team)
- Team pronto was the 1st in final round (summary table below)
  - 2023-08-19/09-01, 360,000 games/team)
  - The winning agent's tactic is not to take others' inappropriate statements seriously, and at the same time make inappropriate statements that mislead others.

Rank	Team	Winnig Rate	Programming Language
1	pronto	63.8%	Java
2	gotsumori	58.2%	Java
3	ddhb	57.1%	Python
4	ioh	56.5%	Java
5	TOT	53.5%	C#

Rank	Team	Winnig Rate	Programming Language
6	cash	52.9%	Python
7	Lunatic	51.4%	Python
8	mikami	49.4%	Java
9	Howls03	49.1%	Java
10	Howls	48.5%	Java
10	MayQueen	48.5%	Java

Natural Language Division

- Communicate via text only (English or Japanese)
  - except for game actions such as vote, attack, divine
- Participant systems played games automatically
  - 5 players (seer + werewolf + possessed + 2 villagers)
  - Self-matches and mutual-matches, each a couple of games
    - Thank Neo Watanabe and Kaito Kagaminuma for hosting games!
- Subjective evaluation performed on the game logs, by human evaluators (organizers), 5 levels for each item below
  - A. **Naturalness**: Did the sentences feel natural?
  - B. **Contextualitly**: Did the sentences feel appropriate to the context, and not a canned answer that can be applied to anything?
  - C. **Consistency in texts**: Did the sentences feel consistent with each other?
  - D. **Consistency with game actions**: Were the game actions (votes, attacks, etc) consistent with the sentences of this agent?
  - E. **Diversity**: Were the sentences interesting? Did the agent feel like having a rich personality?

Game log example by agents

#	Talk	Interpretaion
02	I will call myself a fortune teller. Agent[04] was a human	CO as another seer, divination result to Agent[04] to be a human
01	Agent[02], I am the true fortune teller. Avoid misunderstandings. We should find a werewolf.	Insists that he/she is the true seer
04	>>Agent[02] I see, you guessed that I was human. thank you. But even so, I'm surprised that two fortune tellers came out! Which result should I believe in? What do you guys think?	Consistent response to the divination result of Agent[04]
02	Agent[01] and Agent[02] are said to be fortune tellers, but if there are other fortune tellers among you, please let me know	It is correct that Agent[01] and [02] came out as seers, but [02] is this speaker itself so irrelevant

- # for agent ids. Please refer to our paper for details!

Teams in AIWolf NL 2023

- **5 teams in Japanese, 2 (3) teams in English**
  - **held in June, details in next slide**
- **am** (JP) Mikio Abe, and Akihiro Mikami @ Yamagata University discussion/power play awarded
- **ChatWolf** (JP) Hisaichi Shibata, Soichiro Miki, Yuta Nakamura @ University of Tokyo Hospital
- **HowlGPT** (EN)
- **k2b ara** (JP) Takuya Okubo, Kazuki Takashima, Tomoya Hatanaka, Mami Uchida, Rei Ito @ University of Tokyo character awarded
- **k2b shigemura** (JP) Soga Shigemura, Tomoki Fukuda, Masahiro Wakutani @ Unversity of Tokyo persuation awarded
- **kanolab** (JP/EN) Neo Watanabe, Kaito Kagaminuma, Yoshinobu Kano @ Shizuoka University
- **MIV** (EN) Jaewon Lee and Benedek Hauer @ University of Tokyo
- **sUper_IL** (JP) Zhiyang Qi @ University of Electro-Communications 1st ranked

Results of Subjective Evaluation

- A-E criterion in five levels (as mentioned)
  - Four Japanese organizers for JA
  - Three English fluent staffs for EN
- sUper_IL was the 1st team in total (average)

Team	L	Base Method	Ave.	A	B	C	D	E
am	JA	Rule-based	3.22	3.40	3.35	3.45	**3.80**	2.10
ChatWolf	JA	Original LLM	2.98	3.05	2.40	2.60	2.70	**4.15**
k2b_ara	JA	ChatGPT	3.53	4.07	3.82	3.25	3.07	3.42
k2b_shigemura	JA	ChatGPT	3.35	3.62	3.25	3.12	3.37	3.37
kanolab	JA	ChatGPT	3.77	3.57	3.90	3.75	3.50	**4.15**
sUper_IL	JA	ChatGPT	**4.01**	**4.45**	**4.20**	**4.05**	3.55	3.80
HowlGPT	EN	Various LLMs	2.46	2.66	1.66	2.66	2.00	**3.33**
MIV	EN	ChatGPT	**3.20**	**3.33**	**3.66**	**3.00**	**2.66**	2.33
kanolab	EN	ChatGPT	2.86	3.33	2.66	**3.00**	2.66	2.66

Analysis of Evaluation Results

- ChatGPT (GPT-4) showed good basic performance in sentence generations
  - but differs depending on the prompt and settings
  - Japanese<->English switch was easy by adding an instruction in prompt
- Issues using Generative AIs
  - LLMs other than ChatGPT better in diversity (E)
  - rule-based system was better in consistency with game actions (D)
  - we cannot ask ChatGPT to “tell a lie” directly
  - still difficult to hold “different personalities”
    - i.e. liar and real, other agent thinks me as a liar who would think another agent as a villager and ...
    - exceeds the number of combinations that generative AI could learn?

Summary: Challenging Evaluation for Dialog Systems

- GPT-based LLMs (e.g. ChatGPT) can generate very fluent sentences
  - cannot tell whether automatic generation or not in most cases
  - long context can be coherent, though not perfect
  - now it's time to leap this AIWolf project!
- Shared background knowledge? Deeper Semantics?
  - Coherency with predefined settings and contexts are sometimes broken
  - Contradictory responses suggest that there might be no underlying logic
  - "Aggregations" of LLM instances can solve these issues?
- "Mafia" requires players to carefully understand conversations
  - Evaluations of AIWolf is based on semantic links between game actions and dialogues, rather than just victory/defeat
  - AIWolf contest can reveal the capabilities and limitations of generative AIs

Join in the AIWolf contest 2024!

Questions and Discussions

 # Interactive demo play
with a human (you)
planned in the next poster session

Hamamatsu City, Shizuoka

Self Introduction: Yoshinobu Kano

- Associate Professor (PI) at Faculty of Informatics, Shizuoka University
  - Phd at University of Tokyo (superviser: Prof. Tsujii)
  - JST PRESTO
- Human-like NLP system
  - Spoken language, speech-text integrated parser
  - White-box approach not just end-to-end ML
  - Psychological plausibility
  - Dialog system as an ultimate goal including all human intellectual processing
- Applications on the way of the goal above
  - Question answering, sentence generation, examination solver, legal/medical document processing, automatic diagnosis
- Details at http://kanolab.net/kano/index.en.html

Research Topics

- Human-like NLP/Sentence Generation/Dialog System
  - Sentence generation by psychologically plausible NLP model (MEXT)
  - AI werewolf ("Mafia" game agent) dialog system
  - Ad copy generation (with Dentsu co.)
  - News headline generation (with Chunichi newspapers)
- Medical NLP/Spoken Language
  - Mental disease diagnosis by NLP from conversation and SNS (JST CREST)
  - Developmental disorder (ASD) diagnosis in ADOS by NLP (JST CREST)
  - Side-effect extraction from EHRs
- Text mining for scientific literature in neuroscience (JST CREST)
- Legal NLP/Question Answering
  - Legal Bar Exam Solver in COLIEE (MEXT Kakenhi S)
  - Automatic court judgement
  - Todai Robot to solve university entrance exams
- Public opinion inference and intervention/Political NLP (SECOM, Kakenhi)
  - Prediction and intervention of deceptions in SNS, news articles, perlimament records