Using Graph-Aware Reinforcement Learning to Identify Winning Strategies in Diplomacy Games (Student Abstract)

Hansin Ahuja[1], **Lynnette Hui Xian Ng**[2], **Kokil Jaidka**[3]

[1]Indian Institute of Technology Ropar
[2]Carnegie Mellon University
[3]National University of Singapore
hansinahuja@gmail.com, huixiann@andrew.cmu.edu, jaidka@nus.edu.sg

Abstract

This abstract proposes an approach towards goal-oriented modeling of the detection and modeling complex social phenomena in multiparty discourse in an online political strategy game. We developed a two-tier approach that first encodes sociolinguistic behavior as linguistic features then use reinforcement learning to estimate the advantage afforded to any player. In the first tier, sociolinguistic behavior, such as Friendship and Reasoning, that speakers use to influence others are encoded as linguistic features to identify the persuasive strategies applied by each player in simultaneous two-party dialogues. In the second tier, a reinforcement learning approach is used to estimate a graph-aware reward function to quantify the advantage afforded to each player based on their standing in this multiparty setup. We apply this technique to the game Diplomacy, using a dataset comprising of over 15,000 messages exchanged between 78 users. Our graph-aware approach shows robust performance compared to a context-agnostic setup.

Introduction

In an increasingly connected world, communication for personal, entertainment and professional reasons is increasingly online. However, most research on online communication has focused on personal and professional contexts, while online coordination in game-based contexts is less understood. For example, in multiparty online games, players interact through textual, audio, and visual signals to coordinate game strategies aimed at ultimate victory. This is partly due to the lack of multimodal datasets to examine the dynamic social processes at play.

In this work, we study optimal strategies that interlocutors may employ to maximize chances of success, in a deliberate effort to best other people. This paper examines a recent dataset (Peskov et. al 2020) collected from interacting players during ongoing games of Diplomacy, a political strategy game, as it was played online.

Approach

We adopt a two-tier methodology in predicting the winner of a Diplomacy chat thread. First, we relied on sociolinguistic cues from words and word features to infer persuasive strategies such as Friendship and Reasoning. Next, a reinforcement learning approach was used to construct a graph-aware reward function that considers the in-game dynamic between two players and the interplay of players in a more extensive multiparty setup.

Dataset: The CL-Aff Diplomacy dataset (Jaidka et al. 2021) comprises additional labels to the Diplomacy dataset (Peskov et. al 2020) identifying the rhetorical strategies used by the players in their chat messages. The dataset included four annotated labels about the rhetorical strategies followed by the players (Friendship, Reasoning, Game Move, and Share Information), which had a pairwise inter-annotator agreement of at least 60%. Rather than utterance-level deception, our paper is interested in examining whether signals of influence and persuasion applied in the game's early stages predict ultimate victory.

Weakly Supervised Labeling of Player Strategies

The first step involved identifying the action space for the players in the Diplomacy games in terms of the rhetorical strategies that they follow to influence and persuade each other. However, only 60% of the CL-Aff Diplomacy dataset had high-quality labeled data. Hence, a weakly supervised approach was followed to predict the rhetorical strategies for the entire dataset. The training set comprising the labels from CL-Aff Diplomacy was used to train binary classifiers on the different rhetorical strategies, such as Friendship (F), Reasoning (R), Game Move (GM), and Share Information (SI). The best-performing classifiers (reported in Table 1) were used to predict labels for the entire dataset.[1]

Score-Based Inverse Reinforcement Learning

In the second step, we identified the winning player in each conversation as a function of their rhetorical strategies. We formulate this as an inverse RL problem and, more specifically, use score-based inverse reinforcement learning (SBIRL) (El Asri et al. 2016), which allows us to exploit the experiences of non-expert agents as well.

Each state s of the state space S is encoded using $\phi : S \rightarrow \mathbb{R}^d$. The encoding is done by picking a set of characteristic features corresponding to that state. Simpler operationalizations rely only on the player's score at any given point.

[1] More results can be found at https://tinyurl.com/diplomacy-rl/

On the other hand, a graph-aware operationalization incorporated the player's importance and influence in the textual communication network for the game through several centrality features. For the reward function, a linear parameterization was adopted: $r_\theta(s) = \theta^\intercal \phi(s)$ where θ is the set of parameters. Each thread t of conversation is represented as:

$$t = (s_0, \dots, s_T) = (s_j)_{j=0}^{T}$$

We extracted the thread-level tuples (t_i, f_i), where t_i is the subsequence of states corresponding to the i^{th} player (referred to as a subthread) and f_i is the final score of the i^{th} player at the end of the thread. For any given subthread and reward, the discounted sum of rewards can be written as

$$\sum_{t=0}^{T} \gamma^t r_\theta(s_t) = \theta^\intercal \mu(h) \text{ with } \mu(h) = \sum_{t=0}^{T} \gamma^t \phi(s_t)$$

Where γ is the discounting factor. We regress the final scores f_i on the mappings $\mu(h_i)$ and asymptotically minimise the risk based on the ℓ_2-loss. A reward function estimator r_{θ_n} is derived after estimating θ_n by solving:

$$\theta_n = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (f_i - \theta^\intercal \mu(h_i))^2$$

Preliminary Results

Weakly Supervised Labeling of Player Strategies

Once the annotation labels were obtained, key player strategies were identified. At the player level, Reasoning (R), Game move (GM), and Share Information (SI) all shared a strong pairwise Pearson correlation ($r \in (0.45, 0.55)$), while each were strongly anti-correlated with Friendship (F) ($r \in$ (-0.61, -0.45)), both with $p < 0.001$. This suggests that players either acted on a Friendship or a Reasoning strategy and enabled us to label each utterance as an action under assumptions of mutual exclusion. When both labels were present, a voting mechanism with the Game move and Share Information was used to determine the ultimate label.

	F	**R**	**GM**	**SI**
Logistic regression	0.633	0.508	0.695	0.541
Gaussian NB	0.467	**0.533**	0.620	0.315
Adaboost	0.642	0.448	0.696	**0.623**
Gradient boosting	0.641	0.445	**0.698**	0.611
C-SVC	**0.673**	0.526	0.692	0.570

Table 1: Macro-F1 scores for feature estimators

Efficacy of the Reward Function

The winner of a chat thread is the player with the higher score at the end of the thread. The accuracy of the reward function is the fraction of times when the average estimated reward for the winner was greater than that of the loser.

First, we encoded only the difference between the game scores of the two players in our state feature vector. Next, given that each player is engaged in multiple conversations at the same time, we introduced multiple graph centrality features, such as authority score and eigen centrality, into the state representation[2]. The results are reported in Table 2. We note that our graph-aware approach outperforms the context-agnostic approach, which relies solely on sociolinguistic cues and ignores interplay between players across multiple threads. Additionally, we restrict each player to their first n utterances and perform the same exercise. Fig. 1 shows that the graph-aware approach achieves modest accuracy with the very first half dozen chat messages.

Approach	**Accuracy**
SBIRL	0.71
Graph-aware SBIRL	0.79

Table 2: Reward function accuracy, defined as the fraction of times the winner had a greater average estimated reward

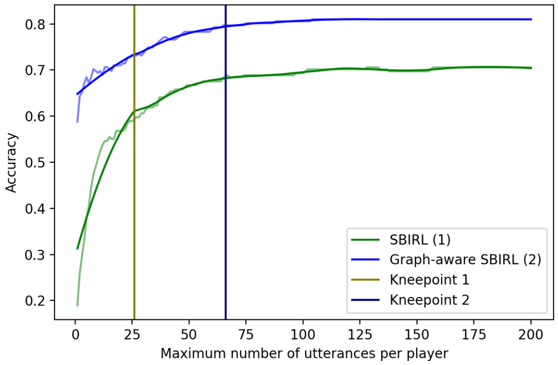

Figure 1: Accuracies with number of utterances restricted

Conclusion and Future Work

The SBIRL approach shows promising results; however, the environment lacks counterfactuals. Judging any one action is difficult because the player never chose the alternate action at that specific point in time, thus making agent-focussed testing difficult. In future experiments, we will explore:

- Complex, non-linear reward function estimators.
- Bootstrapping the dataset to generate counterfactual data, allowing us to build and test well-defined agents.

References

El Asri, L.; Piot, B.; Geist, M.; Laroche, R.; and Pietquin, O. 2016. Score-based inverse reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016)*.

Jaidka, K.; Chhaya, N.; Ungar, L.; Healey, J.; and Sinha, A. 2021. Editorial for the 4th AAAI-21 Workshop on Affective Content Analysis. In *Proceedings of the AAAI-21 Workshop on Affective Content Analysis*. New York, USA: AAAI.

Peskov et. al, D. 2020. It Takes Two to Lie: One to Lie, and One to Listen. In *Proceedings of ACL*.

[2] Calculated using igraph, https://igraph.org/r/