Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning

Tracking Beliefs and Intentions in the Werewolf Game

Codruta Girlea
University of Illinois
Urbana, IL 61801, USA
girlea2@illinois.edu

Eyal Amir
University of Illinois
Urbana, IL 61801, USA
eyal@illinois.edu

Roxana Girju
University of Illinois
Urbana, IL 61801, USA
girju@illinois.edu

Abstract

We propose a model of belief and intention change over the course of a dialogue, in the case where the decisions taken during the dialogue affect the possibly conflicting goals of the agents involved. We use Situation Calculus to model the evolution of the world and an observation model to analyze the evolution of intentions and beliefs. In our formalization, utterances, that only change the beliefs and intentions, are observations. We illustrate our formalization with the game of Werewolf.

Introduction

Agents are engaged in dialogues according to goals they can satisfy by talking to other agents (Cohen and Perrault 1979),(Perrault and Allen 1980). A dialogue is a sequence of utterances, produced by agents based on their beliefs, in order to reach intended states of facts by causing other agents to change their beliefs and thus also their intentions. Our goal is to model this interaction between beliefs, intentions, and utterances. The ability to predict decisions resulting from the dialogue is used as a performance measure.

An example of such a domain is the *Werewolf* game. Here, players are assigned the roles of either villagers or werewolves. The game proceeds in alternating day and night stages, overlooked by an impartial judge. By night, unseen by villagers, the werewolves choose and kill a victim. The victim's identity is announced by the judge at the beginning of the next day. Then, the rest of the villagers discuss and vote to execute one person who they agree is a werewolf. The problem is to predict the outcome of each player's vote.

The game can be modeled with utterances as actions in Situation Calculus (McCarthy 1983), based on Austin's theory of performatives (Austin 1975). A problem is accounting for the indefinitely many perlocutionary acts, or possible effects an utterance may have on the hearer. Belief Revision (Alchourron, Gardenfors, and Makinson 1985) enables modeling and reasoning about changes of beliefs, including as a result of knowledge producing actions in Situation Calculus (Shapiro et al. 2000). Applying Belief Revision to our domain is hard because the agents need to do reasoning about beliefs over beliefs.

We present a new approach to Belief Revision in Situation Calculus. We overcome the need to represent perlocutions by assuming an 'observation model', describing what beliefs, intentions, and unknown properties the utterances expose. The agents' belief states are then filtered (Shirazi and Amir 2011) with the observed utterances, resulting in an updated Kripke structure.

Our new approach allows us to describe dialogues combined with actions. From a Kriple Structure perspective, we add an observation model of utterances, and a revision model, accounting for how the utterances are used to change beliefs. From a Situation Calculus perspective, we model beliefs over beliefs, belief change, and perlocutionary effects. From a Belief Revision perspective, we represent how utterances affect dialogue participants on an individual level, depeding on each agents' internal state.

Our model does not use any linguistic input. We assume the utterances are already parsed to logical formulas that encode both their propositional content and the speech act they function as. Utterances expressed in natural language may provide further insight into the agents' beliefs, encoded in presuppositions and propositional attitudes. We plan to embed those phenomena in our model as future work.

Related Work

Scherl and Levesque (Scherl and Levesque 2003) introduce a situation accessibility relation predicate for the knowledge modality, and a sensing result function for knowledge producing actions, in a single agent scenario. The knowledge of the agent changes only through sensing, and there is no discussion of actions that specifically change the possibly nested beliefs, leaving the world unchanged.

This work has been extended (Shapiro et al. 2000),(Delgrande and Levesque 2012) for Belief Revision and Update. The authors introduce plausabilty as a mapping from possible worlds to natural numbers, defining a total order on possible worlds in which one believes the most plausible possible world. Belief revision changes plausability.

R. Demolombe and P. P. Parra (Demolombe and del Pilar Pozos Parra 2000) define an epistemic logic for the multiagent case, using belief as a simple modality that qualifies sentences but does not allow nesting. Their fragment is tractable, but the fact that it does not allow nested belief makes it inapplicable to domains where agent's actions are

oriented mainly towards changing other agents' beliefs, both about the world and about themselves and other agents.

It is on this (Demolombe and del Pilar Pozos Parra 2000) tractable fragment that most of the work (Demolombe, Mara, and Fern 2005),(Parra, Nayak, and Demolombe 2004) on formalizing intentions or the BDI framework in Situation Calculus is built upon. These formalizations therefore inherit the limited expressing power of the fragment, also not allowing nested beliefs. Here, intentions are about actions, referring to plans, as sequences of actions that are intended, in that specific order, in order to reach the goals. We do not consider planning, our intentions refer to the states of the world currently most desirable to the agent.

None of these approaches can straightforwardly be used in our domain, by enconding utterances as actions, as we need to account for both the change in nested beliefs, and the variability in how an utterance affects the hearer. Early approaches to modeling speech acts as plan operators (Cohen and Perrault 1979) propose a solution to the latter problem by defining mediating acts between the speech acts and the perlocutionary effects. These acts describe the conditions in which the direct effects of the illocutionary act result in the intended effects. This amounts to defining perlocutionary effects as actions where the preconditions are defined on the hearer belief state, and illocutionary effects as actions where the preconditions are defined on the speaker belief state, the latter's effects being carried over to the mediating act's preconditions. A disadvantage is that the set of perlocutionary effects is virtually unbounded for any utterance.

Early work on dialogue modeling (Cohen and Levesque 1985), (Cohen and Levesque 1990) assumes that the dialogue participants are rational, intentional agents that hold beliefs and choose to intend actions according to their goals. This work builds upon earlier theories in pragmatics by Searle (Searle 1976) and Horn (Horn 1984). The authors axiomatize some speech acts as actions in an event calculus, with modal operators for beliefs and goals, and event operators such as *DONE* and *AFTER*. Intention in this framework refers to the intention to execute an action according to a plan in order to reach a goal, so an agent may intend an action, not a set of possible worlds described by a formula.

Formalization of the Observation Model

The observation axioms describe what each observation exposes. The propositional content and function of an utterance is modeled as an observation, rather than an action. The intuition is that utterances do not change the world itself, but change and provide evidence for agent beliefs.

Observation axioms play a similar role to the observation model in statistical models such as the HMM or the POMDP (Kaelbling, Littman, and Cassandra 1998). They also expose a hidden state, using the observed value of a model variable. In our case, the hidden state consists of beliefs, intentions, and unknown properties (roles), and the model variable is a predicate encoding an observable fact.

Preliminary Definitions and Notation

We build upon the framework of Situation Calculus (McCarthy 1983), using successor state axioms as introduced by Reiter (Reiter 1991) as a solution to the frame problem.

We have a set Ag of agents, a set R of roles, and a set S 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), a set M of modalities, a set F of fluents, a set C of constant symbols, and a set A of actions, where: $M = \{K, B, P, I\}$ (knowledge, belief, intention, persistent intention), $F = \{r : \alpha\sigma\}_{r\in R} \cup \{dead : \alpha\sigma, alive : \alpha\sigma, voted : \alpha\alpha\sigma\} \cup \{claim_r : \alpha\alpha\sigma\}_{r\in R} \cup \{claim_{\neg r} : \alpha\alpha\sigma\}_{r\in R}, C = \{c : \alpha\}_{c\in Ag} \cup \{s_0 : \sigma\}, A = \{vote(x)\} \cup \{round_n(x)\}_{3\leq n\leq |C|}$.

Modalities Modalities are ternary relations understood as each agent's accessibility relation on situations (Scherl and Levesque 2003). More specifically, for a modality $M \in \{K, B, P, I\}$, $M(x, s', s)$ 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:

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

To improve readability, we will denote free situation variables with a single mention by symbol _.

Observation Axioms

Utterances are treated as observations, and are produced in the conditions specified by our axiom system.

An observation axiom is an axiom of the form:

$$\forall \vec{x} : \vec{\alpha}.l(\vec{x}, s) \wedge \wedge_{i\neq j\in I} x_i \neq x_j \rightarrow \vee_k(\phi_k \wedge \phi_k^M)$$

where: l is a literal, $\vec{x}$ is a vector of agent variables, I is a subset of the indices of $\vec{x}$, each ϕ_k is a formula with no modalities, and in each ϕ_k^M, every formula with no modalities appears in the scope of a modal formula.

Literal $l(\vec{x}, s)$ is the *observation* and encodes any fact that becomes available to all agents. Formulas ϕ_k encode facts that are not available to all agents. Formulas ϕ_k^M describe the intentions or beliefs that, in the conditions described by ϕ_k, resulted in the observed value of the predicate encoded by l. Formulas ϕ_k and ϕ_k^M describe the hidden state.

We assume only players can make and be the object of claims, for all roles $r \in R$, claims $c \in \{r, \neg r\}$, situations s:

$\forall x, y : \alpha.claim_c(x, y, s) \rightarrow \neg(judge(x, s) \vee judge(y, s))$

As a simplification, we consider as observations claims about the agents' roles. Other utterances only expose their propositional content. The observation model is known by all agents and consists of the following axioms:

- Accusing another player of being the werewolf:

$$\begin{aligned}&\forall x, y : \alpha.(claim_{wolf}(x, y, s) \wedge x \neq y) \rightarrow (wolf(x, s) \wedge \\ &\neg B(x, wolf(y, \_), s) \wedge B(x, B(y, wolf(x, \_))) \vee \\ &\neg wolf(x, s) \wedge B(x, wolf(y, \_), s)) \end{aligned} \quad (1)$$

- Defending another player or oneself:

$$\begin{aligned}&\forall x, y : \alpha.claim_{\neg wolf}(x, y, s) \rightarrow (wolf(x, s) \\ &\wedge B(x, \exists z : \alpha.z \neq x \wedge z \neq y \wedge B(z, wolf(y, \_), \_), s) \\ &\wedge B(x, \exists z : \alpha.z \neq x \wedge z \neq y \wedge \neg B(z, wolf(y, \_), \_), s) \\ &\wedge (x \neq y \vee \neg B(x, wolf(y, \_), s))) \vee (\neg wolf(x, s) \\ &\wedge B(x, \exists z : \alpha.z \neq x \wedge z \neq y \wedge B(z, wolf(y, \_), s), s) \\ &\wedge \neg B(x, wolf(y, \_), s))) \end{aligned} \quad (2)$$

- Claiming the role of seer:
$\forall x : \alpha.claim_{seer}(x, x, s) \rightarrow (wolf(x, s) \vee seer(x, s))$
$\forall x, y : \alpha.(claim_{seer}(x, y, s) \wedge x \neq y) \rightarrow (\neg wolf(x, s)$
$\wedge B(x, seer(y, \_), s) \wedge \neg B(x, wolf(y, \_), s)$
$\wedge B(x, \exists z : \alpha.I(z, dead(y, \_), \_), s))$ (3)

- Claiming that someone is not a seer:
$$\forall x, y : \alpha.claim_{\neg seer}(x, y, s) \rightarrow seer(x, s) \wedge B(x, \exists z : \alpha.B(z, seer(y, \_), s), s) \quad (4)$$

- Not only utterances can be seen as observations that alter beliefs and intentions. Executing an action such as voting can affect fluents such as $voted$, so those fluents can also be modeled as observations:
$$voted(x, y, do(z, a, s)) \rightarrow I(x, dead(y, \_), s) \quad (5)$$

Belief and Intention Axioms

A first set of axioms describe the end goal of each agent, which in the game of Werewolf 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:

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

Initial Situation s_0

Initially, the agents are assigned and thus know their roles: $\forall x : \alpha.r(x, s_0) \rightarrow K(x, r(x, \_), s_0)$ for all roles r.

Since each agent has at most one role, for the unique roles $r \in R_u$ we have:
$$\forall x, y : \alpha.r(x, s) \rightarrow (y = x \vee (K(x, \neg r(y, \_), s_0)) \wedge K(x, \neg B(y, r(y, \_), \_), s_0))) \quad (6)$$

Each game will specify a role assignment. The roles will be assigned in the initial situation, but this assignment will not be available to the agents.

Belief Filtering

Observations are used in the process of logical filtering (Shirazi and Amir 2011) to select the situations to be believed or intended. We keep the original logical filtering notation and definition for formulas not containing modalities.

The result of filtering the belief relation with an observation is a new accessibility relation, where situations inconsistent with the observation are no longer accessible.

Definition 1 *Let M be a modality predicate and o an observation. Then, for any situations s, s' and agent $x \in Ag$:*

$Filter[o][M(x, s', s)] \equiv P(x, s', s)$, *if* $M \in \{P, I\}$

$Filter[o][M(x, s', s)] \equiv o_{[s/s']} \wedge M(x, s', s)$, *otherwise*

An agent will believe the consequences of her previous beliefs and the new observation:

Theorem 1 *For any modality $M \in \{K, B\}$, agent Ag, situation s, formulas ϕ, ψ, and observation o with $\{\psi, o\} \models \phi$:*

$$\{M(x, \psi, s)\} \models Filter[o][M(x, \phi, s)]$$

Conversely, only those consequences of previous beliefs that are consistent with the new observation are believed:

Theorem 2 *For any modality $M \in \{K, B\}$, agent x, situation s, formulas ϕ, ψ, and observation o with $\{\psi, \phi\} \models \neg o$:*

$$\{M(x, \psi, s), Filter[o][M(x, \phi, s)]\} \models \Box$$

This does not hold for intentions: as a result of refining her beliefs, an agent can drastically change her intentions.

Formalization of the Dynamic Model

We model actions that affect the world in Situation Calculus. The actions are voting and ending a round of game.

Voting is available to any player and any living player can vote for any living player she intends the death of:
$$Poss(x, vote(y), s) \equiv \neg judge(x, s) \wedge \neg judge(y, s) \wedge I(x, dead(y, \_), s) \quad (7)$$

Ending a game round of n players with x voted out ($round_n(x)$) is available to the judge:

$Poss(u, round_n(x), s) \equiv$
$\exists y_1 ... y_n : \alpha.(\wedge_{1 \leq i \leq n}(alive(y_i, \_) \wedge \neg judge(y_i, \_)) \wedge$
$\wedge_{1 \leq i \neq j \leq n} y_i \neq y_j \wedge (\forall z : \alpha.alive(z, \_) \wedge \neg judge(z, \_) \rightarrow$
$\vee_{1 \leq i \leq n} z = y_i)) \wedge judge(u, s) \wedge \neg dead(x, s) \wedge$
$\exists y_1 ... y_{1+n/2} : \alpha. \wedge_{1 \leq i \leq 1+n/2} voted(y_i, x, s)$

Ending a game round is also an *announcement* action, where the judge announces the identity of the player who was killed by werewolves during the night. Similarly to a sensing action (Scherl and Levesque 2003), it has associated an *announcement result* AR. This is a partial function as, depending on the roles, there may be rounds with no victims.

$AR(round_n(x), s, r) \equiv \vee_{c \in Ag}(r = c \wedge dead(c, s))$
$AR(round_n(x), s, r) \wedge AR(round_n(x), s, r') \rightarrow r = r'$

Successor State Axioms

The successor state axioms describe the effects of executing an action on any of the fluents:

- The roles do not change:
$r(x, do(y, a, s)) \equiv r(x, s) \wedge \neg dead(x, do(y, a, s))$
- Players are dead because they were either already dead, voted out at the end of the current round, or killed by werewolves at the end of the current round:
$dead(x, do(z, a, s)) \equiv dead(x, s) \vee (a = round_n(x) \wedge Poss(z, a, s)) \vee \exists y : \alpha.(a = round_n(y) \wedge Poss(z, a, s) \wedge AR(a, do(z, a, s), x))$

- A player's vote is counted when she performs the action of voting and does not change unless she votes for another player:

$$voted(x, y, do(u, a, s)) \equiv$$
$$(voted(x, y, s) \wedge (u \neq x \vee \forall z : \alpha.a \neq vote(z)))$$
$$\vee (a = vote(y) \wedge u = x \wedge Poss(x, a, s))$$

Note that belief filtering and action execution are orthogonal, in that the first only affects accessibility relations, whereas the latter only affects the current situation. Action execution results in a new situation, and new observations will lead to refining the accessibility of situations believed or intended from the new situation.

Conclusions

We presented a model for beliefs in dialogues where agents have conflicting intentions, and the Werewolf game in particular. Other domains can be formalized in a similar manner, by writing the appropriate observation model and belief and intention interaction axioms.

A possible line of future work is using the formalism for a richer domain, such as trials. Another extension is to probabilistic observation and dynamic models, and probabilistic belief and intention accessibility relations. This would allow better predictions of the voting outcomes, as an agent will no longer choose an action non-deterministically, but will have a preference over intended outcomes.

This formalization could be useful in computational pragmatics, for modeling complex conversations where some reasoning beyond text level is needed to follow and predict what the participants may say. In turn, linguistic cues exposing attitudes and polarity could result in a more accurate model of human interaction. Furthermore, in our example, we assumed the dialogue already parsed in a sequence of observations and relations. In the future, we would need to extract these logical formulas from text.

Acknowledgments

This work was supported by NSF IIS grants 09-17123, 09-68552, grant NSF EAR 09-43627 EA, and a Defense Advanced Research Project Agency (DARPA) grant as part of the Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA, AFRL, or the US government.

References

Alchourron, C. E.; Gardenfors, P.; and Makinson, D. 1985. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic* 50:510–530.

Austin, J. L. 1975. *How to do things with words*. Cambridge, Mass.: Harvard University Press.

Cohen, P. R., and Levesque, H. J. 1985. Speech acts and rationality. In Mann, W. C., ed., *ACL*, 49–60. ACL.

Cohen, P. R., and Levesque, H. J. 1990. Performatives in a rationally based speech act theory. In Berwick, R. C., ed., *ACL*, 79–88. ACL.

Cohen, P. R., and Perrault, C. R. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science* 3(3):177–212.

Delgrande, J. P., and Levesque, H. J. 2012. Belief revision with sensing and fallible actions. In Brewka, G.; Eiter, T.; and McIlraith, S. A., eds., *KR*. AAAI Press.

Demolombe, R., and del Pilar Pozos Parra, M. 2000. A simple and tractable extension of situation calculus to epistemic logic. In Ras, Z. W., and Ohsuga, S., eds., *ISMIS*, volume 1932 of *Lecture Notes in Computer Science*, 515–524. Springer.

Demolombe, R.; Mara, A.; and Fern, O. 2005. Intention recognition in the situation calculus and probability theory frameworks. In *In Computational Logic in Multi Agent Systems*, 358–372.

Horn, L. 1984. Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. *1984* 1142. Cited by 615.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artif. Intell.* 101(1-2):99–134.

McCarthy, J. 1983. Situations, actions, and causal laws. Technical Report Memo 2, Stanford Artificial Intelligence Project, Stanford University.

Parra, P. P.; Nayak, A.; and Demolombe, R. 2004. Theories of intentions in the framework of situation calculus. In *Declarative Agent Languages and Technologies (DALT 2004), LNCS 3476, Springer-Verlag (2005). In this volume*. Springer Verlag.

Perrault, C. R., and Allen, J. F. 1980. A plan-based analysis of indirect speech acts. *Comput. Linguist.* 6(3-4):167–182.

Reiter, R. 1991. Artificial intelligence and mathematical theory of computation. San Diego, CA, USA: Academic Press Professional, Inc. chapter The Frame Problem in Situation the Calculus: A Simple Solution (Sometimes) and a Completeness Result for Goal Regression, 359–380.

Scherl, R., and Levesque, H. J. 2003. Knowledge, action, and the frame problem. *Artificial Intelligence* 144(1–2):1–39.

Searle, J. R. 1976. A classification of illocutionary acts. *Language in Society* 5:1–23.

Shapiro, S.; Pagnucco, M.; Lesprance, Y.; and Levesque, H. J. 2000. Iterated belief change in the situation calculus. In *Principles of Knowledge Rep. and Reasoning: Proc. of the 7th Int. Conf*, 527–538.

Shirazi, A., and Amir, E. 2011. First-order logical filtering. *Artif. Intell.* 175(1):193–219.