An Answer Set Programming Framework for Reasoning about Truthfulness of Statements by Agents

Tran Cao Son[1], Enrico Pontelli[1], Michael Gelfond[2], and Marcello Balduccini[3]

1 **Dept. Computer Science, New Mexico State University, Las Cruces, NM, USA**
`tson|epontell@cs.nmsu.edu`
2 **Dept. Computer Science, Texas Tech University, Lubbock, TX, USA**
`michael.gelfond@ttu.edu`
2 **Dept. Computer Science, Drexel University, Philadelphia, PA, USA**
`marcello.balduccini@drexel.edu`

Abstract

We propose a framework for answering the question of whether statements made by an agent can be believed, in light of observations made over time. The basic components of the framework are a formalism for reasoning about actions, changes, and observations and a formalism for default reasoning. The framework is suitable for concrete implementation, e.g., using answer set programming for asserting the truthfulness of statements made by agents, starting from observations, knowledge about the actions of the agents, and a theory about the "normal" behavior of agents.

1998 ACM Subject Classification F.4.1, I.2.3, I.2.4, I.2.11

Keywords and phrases Agents, ASP, Reasoning, Knowledge

Digital Object Identifier 10.4230/OASIcs..2016.NNN

1 Introduction

In this extended abstract, we are interested in reasoning about the truthfulness of statements made by agents. We assume that we can observe the world as well as agents' actions. The basis for our judgments will be composed of our observations, performed along a linear time line, along with our commonsense knowledge about agents' behavior and the world. We assume that observations are *true* at the time they are made, and will stay true until additional pieces of information prove otherwise. Our judgments reflect what we believe. They might not correspond to the ground truth and could change over time. This is because we often have to make our judgment in presence of incomplete information. This makes reasoning about the truthfulness of statements made by agents *non-monotonic*. Furthermore, our judgment against a statement is *independent* of whether or not we trust the agent from whom the statement originated. This is illustrated in the next example.

▶ **Example 1.**
- *Time* t_0: When we first meet, John said that his family is poor (property *poor* is true). It is likely that we would believe John—since we have no reasons to conclude otherwise.
- *Time* t_1: We observe the fact that John attends an expensive college (property *in_college* is true). Since students attending the college are *normally* from rich families (default d_1), this would lead us to conclude that John has lied to us. We indicate that the default d_1 is the reason to draw such conclusion, i.e., we changed our belief on the property *poor*.
- *Time* t_2: We observe the fact that John has a need-based scholarship (property *has_scholarship* is true). Since a student's hardship is *usually* derived from the family's financial situation (default

Technical Communications of the 32nd Int'l Conference on Logic Programming (ICLP 2016).
Editors: Manuel Carro, Andy King, Marina De Vos, and Neda Saeedloei; Article No. NNN; pp. NNN:1–NNN:4
Open Access Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

d_2), this fact allows us to withdraw the conclusion that John is a liar, made at time instance t_1. It is still insufficient for us to conclude that John's family is poor.
The situation might be different if, for example, we have a preference among defaults. In this example, if we are inclined to believe in the conclusion of d_2 more than that of d_1, then we would believe that John's family is poor and thus restore our trust in John's original statement (i.e., truth of *poor*).

In this extended abstract, we

1. present the formalization of an abstract model to represent and reason about truthfulness of agent's statements (briefly summarized in the next section); and
2. discuss the steps for a concrete realization of the model using Answer Set Programming.

2 A General Model for Reasoning about Truthfulness of Statements made by Agents

In this section, we propose a general framework for representing, and reasoning about, the truthfulness of (statements made by) agents[1]. The framework can be instantiated using specific paradigms for reasoning about actions and change and for non-monotonic reasoning. We assume that

- It is possible to observe the properties of the world and the occurrences of the agents' actions over time (e.g., we observe that John buys a car, John is a student, etc.). Let us denote with O_a and O_w the set of action occurrences and the set of observations about the world over time, respectively.
- We have adequate knowledge about the agents' actions and their effects (e.g., the action of buying a car requires that the agent has money and its execution will result in the agent owning a car). This knowledge is represented by an action theory Act in a suitable logic A, that allows reasoning about actions' effects and consequent changes to the world. Let $\models_A$ denote the entailment relation defined within the logical framework A used to describe Act.
- We have commonsense knowledge about "normal" behavior (e.g., a person attending an expensive school normally comes from a rich family, a person obtaining need-based scholarship usually comes from a poor family). This knowledge is represented by a default theory with preferences Def, that enables reasoning about the state of the world and deriving conclusions whenever necessary. Let $\models_D$ denote the entailment relation defined over the default theory framework defining Def.

The set of observations O_w in Example 1 includes the observations such as 'John comes from a poor family' at time point t_0, 'John attends an expensive college' at time point t_1, and 'John receives a need-based scholarship' at time point t_2. In this particular example we do not have any action occurrences, i.e., $O_a = \emptyset$. Our default theory D consists of d_1 and d_2, which allow us to make conclusions regarding whether John comes from a rich family or not.

Let us consider a theory $T = (O_a, O_w, Act, Def)$ and the associated entailment relations $\models_A$ and $\models_D$. We are interested in answering the question of whether a statement asserting that a proposition p is true at a certain time step t, denoted by $p[t]$, is true or not. Specifically, we would like to define the entailment relation $\models$ between T and $p[t]$. Intuitively, this can be done in two steps:

- Compute possible models $W[t]$ of the world at the time point t from Act, O_a, and O_w ($\models_A$); and
- Determine whether p is true given Def and $W[t]$ (using $\models_D$).

[1] From now on, we will often use "the truthfulness of agents" interchangeably with "the truthfulness of statements made by agents."

Let us assume that the equation $W[t] = \{z \mid Act \cup O_a \cup O_w \models_A z[t]\}$ characterizes any of the states of the world at time step t given Act, O_a, and O_w (based on the semantics of $\models_A$). The entailment relation between T and $p[t]$ can be defined as follows.

$$T \models p[t] \quad \Leftrightarrow \quad \forall W[t].\left(\ W[t] = \{z \mid Act \cup O_a \cup O_w \models_A z[t]\} \Rightarrow Def \cup W[t] \models_D p\ \right) \tag{1}$$

Note that this definition also allows one to identify elements of O_a and O_w which, when obtained, will result in the confirmation or denial of $T \models p[t]$. As such, a system that obeys (1) can also be used by users who are interested in what they need to do in order to believe in a statement about p at the time step t, given their beliefs about the behavior of the observed agents.

3 Reasoning about Truthfulness of Agents Using ASP

To develop a concrete system for reasoning about truthfulness of agents using (1), specific formalizations of *Act* and *Def* need to be developed. There is a large body of research related to these two areas, and deciding which one to use depends on the system developer. Well-known formalisms for reasoning about actions and change, such as action languages [4], situation calculus [8], etc., can be employed for *Act* (and $\models_A$). Approaches to default reasoning with preferences, such as those proposed in [1, 2, 3, 5]), can be used for *Def* (and $\models_D$). In addition, let us note that, in the literature, $\models_D$ can represent *skeptical* or *credulous* reasoning; and the model does not specify how observations are collected. Deciding which type of reasoning is suitable or how to collect observations is an important issue, but it is application-dependent and beyond the scope of this extended abstract. Using the formalisms in [5] and [4] for default reasoning and reasoning about actions and change, respectively, we can implement a system for reasoning about truthfulness of agents using answer set programming (ASP) [6, 7] with the following steps:

- Extending the framework in [5] to allow observations at different time points and developing ASP rules for reasoning with observations; for instance, the language needs to allow facts of the form $obs(p,t)$ —fluent literal p is true at time t—and ASP rules for reasoning about defaults and rules given observations at different time point need to be developed.
- Defining a query language for reasoning about statements of agents at different time points; more specifically, given an ASP program Π encoding the theory described in the previous item and a statement $stm(p,t)$—stating that literal p holds at time t—how does Π helps identify whether or not the statement is true or false; for instance, one can say that if Π entails $p[t]$ with respect to the answer set semantics then the statement is true.
- Allowing observations of the form $occ(a,t)$—action a occurs at time t—and developing ASP rules for reasoning about preconditions of actions as well as effects of actions need to be included. More specifically, we can add ASP rules stating that if an action a occurs at time point t then its preconditions must hold at time t, i.e., its preconditions must be observed at time t; furthermore, its effects must hold (or be observed) at time $t+1$.

4 Conclusions

We proposed a general framework for reasoning about the truthfulness of statements made by an agent. We discussed how the framework can be implemented using ASP using well-known methodologies for reasoning about actions and change and for default reasoning with preferences. The framework does not assume complete knowledge about the agent being observed and the reasoning process builds on observations about the state of the world and occurrences of actions. We had developed an ASP implementation of the framework and explored the use of the framework in simple scenarios derived from man-in-the-middle attacks. The details can be found in the full version of this extended abstract.

References

1 G. Brewka and T. Eiter. Preferred answer sets for extended logic programs. *Artificial Intelligence*, 109:297–356, 1999.

2 G. Brewka and T. Eiter. Prioritizing default logic. In *Intellectics and Computational Logic*, volume 19 of *Applied Logic Series*, pages 27–45. Kluwer, 2000.

3 J. Delgrande, T. Schaub, and H. Tompits. A framework for compiling preferences in logic programs. *Theory and Practice of Logic Programming*, 3(2):129–187, March 2003.

4 M. Gelfond and V. Lifschitz. Action Languages. *Electronic Transactions on Artificial Intelligence*, 3(6), 1998.

5 M. Gelfond and T. C. Son. Reasoning about prioritized defaults. In *Selected Papers from the Workshop on Logic Programming and Knowledge Representation 1997*, pages 164–223. Springer Verlag, LNAI 1471, 1998.

6 V. Marek and M. Truszczyński. Stable models and an alternative logic programming paradigm. In *The Logic Programming Paradigm: a 25-year Perspective*, pages 375–398, 1999.

7 I. Niemelä. Logic programming with stable model semantics as a constraint programming paradigm. *Annals of Mathematics and Artificial Intelligence*, 25(3,4):241–273, 1999.

8 R. Reiter. *KNOWLEDGE IN ACTION: Logical Foundations for Specifying and Implementing Dynamical Systems*. The MIT Press, 2001.