Model Reconciliation in Logic Programs — Extended Abstract

Tran Cao Son\textsuperscript{1} and Van Nguyen\textsuperscript{1} Stylianos Loukas Vasileiou\textsuperscript{2} and William Yeoh\textsuperscript{2}

\textsuperscript{1}New Mexico State University, Las Cruces, NM 88003, USA
{tson, vnguyen}@cs.nmsu.edu
\textsuperscript{2}Washington University in St. Louis, Saint Louis, MO 63130, USA
{v.stylianos, wyeoh}@wustl.edu

Introduction. The \textit{model reconciliation problem} (MRP) has been introduced and investigated in the context of planning \cite{12456}, where one agent (e.g., a robot) needs to explain to another agent (e.g., a human) the question \textit{“why a certain plan is an optimal plan?”}. MRP defines the notion of an explanation from an agent to a human as a pair \((\epsilon^+, \epsilon^-)\) of sets where \(\epsilon^-\) and \(\epsilon^+\) is a set of information, such as preconditions or post-conditions of actions, that the human should add and remove from their problem definition, respectively. The focus has been on developing algorithms for computing explanations that are optimal with respect to some criteria (e.g., the minimality of \(|\epsilon^+ \cup \epsilon^-|\)).

This is an extended abstract of our JELIA paper \cite{3}. It proposes a generalization of MRP in the context of answer set programming and defines the \textit{model reconciliation problem in logic programs} (MRLP). Given \(\pi_a\) and \(\pi_h\), which represent the knowledge bases of an agent and a human, respectively, and a query \(q\) such that \(\pi_a\) entails \(q\) and \(\pi_h\) does not entail \(q\) (or \(\pi_a\) does not entail \(q\) and \(\pi_h\) entails \(q\)), MRLP focuses on determining a pair \((\epsilon^+, \epsilon^-)\) such that \(\epsilon^+ \subseteq \pi_a\), \(\epsilon^- \subseteq \pi_h\), and the program \(\hat{\pi}_h = (\pi_h \backslash \epsilon^-) \cup \epsilon^+\) has an answer set containing \(q\) (or has no answer set containing \(q\)). The pair \((\epsilon^+, \epsilon^-)\) is referred to as a \textit{solution} for the model reconciliation problem \((\pi_a, \pi_h, q)\) (or \((\pi_a, \pi_h, \neg q)\)). The paper discusses different characterizations of solutions and algorithms for computing solutions for MRLP.

Model Reconciliation in Logic Programs (MRLP). A general MRLP is defined as a combination of two sub-problems: One aims at changing the human program so that it entails an atom (e-MRLP) and another focuses on achieving that the updated program does not entail an atom (n-MRLP). Let \(\pi_a\) and \(\pi_h\) be two logic programs and \(q\) be an atom in the language of \(\pi_a\).

- The problem of \textit{model reconciliation for entailment in logic programs} (e-MRLP) is defined by a triple \((\pi_a, \pi_h, q)\). A pair of programs \((\epsilon^+, \epsilon^-)\), such that \(\epsilon^- \subseteq \pi_a\) and \(\epsilon^- \subseteq \pi_h\) is a solution of \((\pi_a, \pi_h, q)\) if \(\hat{\pi}_h \models q\), where \(\hat{\pi}_h = \pi_h \backslash \epsilon^- \cup \epsilon^+\).

- The problem of \textit{model reconciliation for non-entailment in logic programs} (n-MRLP) is defined by a triple \((\pi_a, \pi_h, \neg q)\). A pair of programs \((\epsilon^+, \epsilon^-)\), such that \(\epsilon^+ \subseteq \pi_a\) and \(\epsilon^- \subseteq \pi_h\) is a solution of \((\pi_a, \pi_h, \neg q)\) if \(\hat{\pi}_h \not\models q\), where \(\hat{\pi}_h = \pi_h \backslash \epsilon^- \cup \epsilon^+\).

- The general problem of \textit{model reconciliation in logic programs} (MRLP) is defined by a triple \((\pi_a, \pi_h, \omega)\) where \(\omega = \omega^+ \land \neg \omega^-\) and \(\omega^+\) (resp. \(\omega^-\)) is a conjunction of atoms in \(\pi_a\). \((\epsilon^+, \epsilon^-)\) is a solution for the MRLP problem if it is a solution for \((\pi_a, \pi_h, q)\) for each conjunct \(q\) in \(\omega^+\) and solution for \((\pi_a, \pi_h, \neg r)\) for each conjunct \(r\) in \(\omega^-\).

---

\textsuperscript{1}This research is partially supported by NSF grants 1757207, 1812619, 1812628, and 1914635.

\textsuperscript{2} We say a program \(\pi\) entails (resp. does not entail) an atom \(q\), denoted by \(\pi \models q\) (resp. \(\pi \not\models q\)), if \(q\) belongs to an answer set of \(\pi\) (resp. does not belong to any answer set of \(\pi\)).

\textsuperscript{*}© T. C. Son, V. Nguyen, S. L. Vasileiou, W. Yeoh
This work is licensed under the Creative Commons Attribution License.
Characterizing Solutions. A MRLP might have several solutions and choosing a suitable solution is application-dependent. Some characteristics of solutions that could influence the choice are defined next. Let \((\pi_a, \pi_h, \omega)\) be an MRLP problem and \((\epsilon^+, \epsilon^-)\) be a solution of \((\pi_a, \pi_h, \omega)\). We say:

- \((\epsilon^+, \epsilon^-)\) is **optimal** if there exists no solution \((\lambda^+, \lambda^-)\) such that \(\lambda^+ \cup \lambda^- \subseteq \epsilon^+ \cup \epsilon^-\).
- \((\epsilon^+, \epsilon^-)\) is **\(\pi\)-restrictive** for \(\pi \subseteq \pi_a\) if \(\epsilon^+ \subseteq \pi\); it is **minimally-restrictive** if there exists no solution \((\lambda^+, \lambda^-)\) such that \(\lambda^+ \subseteq \epsilon^+\).
- \((\epsilon^+, \epsilon^-)\) is **\(\pi\)-preserving** for \(\pi \subseteq \pi_h\) if \(\pi \cap \epsilon^- = \emptyset\); it is **maximally-preserving** if there exists no solution \((\lambda^+, \lambda^-)\) such that \(\lambda^- \subseteq \epsilon^-\).
- \((\epsilon^+, \epsilon^-)\) is **assertive** if every answer set of \(\pi_h \setminus \epsilon^- \cup \epsilon^+\) satisfies \(\omega\).
- \((\epsilon^+, \epsilon^-)\) is a **solution with justification** (or \(j\)-solution) if \(\epsilon^+\) contains a justification for \(\omega^+\) w.r.t. some answer set \(I\) of \(\pi_a\).
- \((\epsilon^+, \epsilon^-)\) is a **fact-only** if \(\epsilon^+\) and \(\epsilon^-\) are set of facts in \(\pi_a\) and \(\pi_h\), respectively.

Each class of solutions has its own merits and could be useful in different situations. Optimal solutions could be useful when solutions are associated with some costs.\(^\text{2}\) Minimally-restrictive solutions focus on minimizing the amount of information that the agent needs to introduce to the human and could be useful when explaining a new rule is expensive. On the other hand, maximally-preserving solutions are appropriate when one seeks to minimize the amount of information that needs to be removed from the human knowledge base. Solutions with justifications are those that come with their own support. Assertive solutions do not leave the human any reason for questioning the formula in discussion. Fact-only solutions are special in that they inform the human of their missing or false facts. As a planning problem can be encoded by a logic program whose answer sets encode solutions of the original planning problem, it is easy to see that an MRP in planning can be encoded as an MRLP whose fact-only solutions encode the solutions of the original MRP.

Computing Solutions. Let \(\pi_a\) and \(\pi_h\) be two programs and \(I\) be a set of atoms of \(\pi_a\) and \(\epsilon^+ \subseteq \pi_a\). \(\otimes(\pi_h, \epsilon^+, I)\) is the collection of rules from \(\pi_h \setminus \epsilon^+\) such that for each rule \(r \in \otimes(\pi_h, \epsilon^+, I)\):

1. **head**(\(r\)) \(\in I\)
2. **neg**(\(r\)) \(\cap I = \emptyset\)
3. **pos**(\(r\)) \(\setminus I \neq \emptyset\)

Let \(\epsilon^-[\epsilon^+, I, \pi_a]\) denote the set of rules \(\pi_h \setminus \otimes(\pi_h, \epsilon^+, I)\). This can be used for computing solutions of general MRLP problems as follows. Without loss of generality, consider the problem \((\pi_a, \pi_h, q \land \neg r)\), where \(q\) and \(r\) are atoms of \(\pi_a\) and \(\pi_h\) \(\vdash q\) and \(\pi_a \not\vdash r\). A solution \((\epsilon^+, \epsilon^-)\) for \((\pi_a, \pi_h, q \land \neg r)\) can be computed by the following steps:

1. Compute an answer set \(I\) of atoms that supports \(q\) and identify a minimal justification \(\epsilon^+\) of \(q\) w.r.t. \(I\).
2. Compute \(\epsilon^- = \epsilon^-[\epsilon^+, I, \pi_a]\).
3. Identify a set of rules \(\lambda\) from \(\pi_a \setminus \epsilon \cup \epsilon^+\) such that \(\pi_a \not\vdash \lambda\).

The final solution for \((\pi_a, \pi_h, q \land \neg r)\) is then \((\epsilon^+, \epsilon^- \cup \lambda)\). This process can be implemented using answer set programming.

Conclusions and Future Work. The paper discusses the MRLP problem and its theoretical foundation such as the definition of a solution, the classification of solutions, or methods for computing solutions. The present work assumes that the agent, who needs to compute solutions, has the knowledge of both programs \(\pi_a\) and \(\pi_h\). In practice, this assumption is likely invalid and the agent might also needs to change its program through communication or dialogue with the human. Addressing this issue and developing a system for computing solutions of MRLPs are our immediate future work.

\(^\text{2}\)By associating costs to rules or atoms (e.g., via a cost function), the cost of a solution \((\epsilon^+, \epsilon^-)\) can be defined and used as a criteria to evaluate solutions.
References


