Model Reconciliation in Logic Programs — Extended Abstract*

Tran Cao Son[1] and Van Nguyen[1] Stylianos Loukas Vasileiou[2] and William Yeoh[2]
[1]New Mexico State University, Las Cruces, NM 88003, USA
{tson,vnguyen}@cs.nmsu.edu
[2]Washington University in St. Louis, Saint Louis, MO 63130, USA
{v.stylianos,wyeoh}@wustl.edu

Introduction. The *model reconciliation problem* (MRP) has been introduced and investigated in the context of planning [1, 2, 4, 5, 6], where one agent (e.g., a robot) needs to explain to another agent (e.g., a human) the question "*why a certain plan is an optimal plan?*" MRP defines the notion of an explanation from an agent to a human as a pair $(\varepsilon^+, \varepsilon^-)$ of sets where ε^- and ε^+ 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 $|\varepsilon^+ \cup \varepsilon^-|$).

This is an extended abstract of our JELIA paper [3]. It proposes a generalization of MRP in the context of answer set programming and defines the *model reconciliation problem in logic programs* (MRLP). Given π_a and π_h, which represent the knowledge bases of an agent and a human, respectively, and a query q such that π_a entails[1] q and π_h does not entail q (or π_a does not entail q and π_h entails q), MRLP focuses on determining a pair $(\varepsilon^+, \varepsilon^-)$ such that $\varepsilon^+ \subseteq \pi_a$, $\varepsilon^- \subseteq \pi_h$, and the program $\hat{\pi}_h = (\pi_h \setminus \varepsilon^-) \cup \varepsilon^+$ has an answer set containing q (or has no answer set containing q). The pair $(\varepsilon^+, \varepsilon^-)$ is referred to as a *solution* for the model reconciliation problem (π_a, π_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 π_a and π_h be two logic programs and q be an atom in the language of π_a.

- The problem of *model reconciliation for entailment in logic programs* (e-MRLP) is defined by a triple (π_a, π_h, q). A pair of programs $(\varepsilon^+, \varepsilon^-)$, such that $\varepsilon^+ \subseteq \pi_a$ and $\varepsilon^- \subseteq \pi_h$ is a *solution* of (π_a, π_h, q) if $\hat{\pi}_h \mid\sim q$, where $\hat{\pi}_h = \pi_h \setminus \varepsilon^- \cup \varepsilon^+$.
- The problem of *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 $(\varepsilon^+, \varepsilon^-)$, such that $\varepsilon^+ \subseteq \pi_a$ and $\varepsilon^- \subseteq \pi_h$ is a *solution* of $(\pi_a, \pi_h, \neg q)$ if $\hat{\pi}_h \not\mid\sim q$, where $\hat{\pi}_h = \pi_h \setminus \varepsilon^- \cup \varepsilon^+$.
- The general problem of *model reconciliation in logic programs* (MRLP) is defined by a triple (π_a, π_h, ω) where $\omega = \omega^+ \wedge \neg\omega^-$ and ω^+ (resp. ω^-) is a conjunction of atoms in π_a. $(\varepsilon^+, \varepsilon^-)$ is a solution for the MRLP problem if it is a solution for (π_a, π_h, q) for each conjunct q in ω^+ and solution for $(\pi_a, \pi_h, \neg r)$ for each conjunct r in ω^-.

*This research is partially supported by NSF grants 1757207, 1812619, 1812628, and 1914635.

[1] We say a program π entails (resp. does not entail) an atom q, denoted by $\pi \mid\sim q$ (resp. $\pi \not\mid\sim q$), if q belongs to an answer set of π (resp. does not belong to any answer set of π).

Submitted to:
ICLP 2022

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 (π_a, π_h, ω) be an MRLP problem and $(\varepsilon^+, \varepsilon^-)$ be a solution of (π_a, π_h, ω). We say:

- $(\varepsilon^+, \varepsilon^-)$ is *optimal* if there exists no solution (λ^+, λ^-) such that $\lambda^+ \cup \lambda^- \subset \varepsilon^+ \cup \varepsilon^-$.
- $(\varepsilon^+, \varepsilon^-)$ is π*-restrictive* for $\pi \subseteq \pi_a$ if $\varepsilon^+ \subseteq \pi$; it is *minimally-restrictive* if there exists no solution (λ^+, λ^-) such that $\lambda^+ \subset \varepsilon^+$.
- $(\varepsilon^+, \varepsilon^-)$ is π*-preserving* for $\pi \subseteq \pi_h$ if $\pi \cap \varepsilon^- = \emptyset$; it is *maximally-preserving* if there exists no solution (λ^+, λ^-) such that $\lambda^- \subset \varepsilon^-$.
- $(\varepsilon^+, \varepsilon^-)$ is *assertive* if every answer set of $\pi_h \setminus \varepsilon^- \cup \varepsilon^+$ satisfies ω.
- $(\varepsilon^+, \varepsilon^-)$ is a *solution with justification* (or *j-solution*) if ε^+ contains a justification for ω^+ w.r.t. some answer set I of π_a.
- $(\varepsilon^+, \varepsilon^-)$ is a *fact-only* if ε^+ and ε^- are set of facts in π_a and π_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.[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 π_a and π_h be two programs and I be a set of atoms of π_a and $\varepsilon^+ \subseteq \pi_a$. $\otimes(\pi_h, \varepsilon^+, I)$ is the collection of rules from $\pi_h \setminus \varepsilon^+$ such that for each rule $r \in \otimes(\pi_h, \varepsilon^+, I)$: (*i*) $head(r) \in I$ and $neg(r) \cap I = \emptyset$; or (*ii*) $neg(r) \cap heads(\varepsilon^+) \neq \emptyset$; or (*iii*) $pos(r) \setminus I \neq \emptyset$. Let $\varepsilon^-[\varepsilon^+, I, \pi_h]$ denote the set of rules $\pi_h \setminus \otimes(\pi_h, \varepsilon^+, I)$. This can be used for computing solutions of general MRLP problems as follows. Without loss of generality, consider the problem $(\pi_a, \pi_h, q \wedge \neg r)$, where q and r are atoms of π_a and $\pi_a \mid\sim q$ and $\pi_a \mid\not\sim r$. A solution $(\varepsilon^+, \varepsilon^-)$ for $(\pi_a, \pi_h, q \wedge \neg r)$ can be computed by the following steps: (*i*) compute an answer set I of π_a that supports q and identify a minimal justification ε^+ of q w.r.t. I; (*ii*) compute $\varepsilon^- = \varepsilon^-[\varepsilon^+, I, \pi_h]$; and (*iii*) identify a set of rules λ from $\pi' = \pi_h \setminus \varepsilon \cup \varepsilon^+$ so that $\pi' \setminus \lambda \mid\not\sim r$. The final solution for $(\pi_a, \pi_h, q \wedge \neg r)$ is then $(\varepsilon^+, \varepsilon^- \cup \lambda)$. 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 π_a and π_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.

[2] By associating costs to rules or atoms (e.g., via a cost function), the cost of a solution $(\varepsilon^+, \varepsilon^-)$ can be defined and used as a criteria to evaluate solutions.

References

[1] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang & Subbarao Kambhampati (2017): *Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy*. In: *IJCAI*, pp. 156–163, doi:10.24963/ijcai.2017/23.

[2] Van Nguyen, Stylianos Loukas Vasileiou, Tran Cao Son & William Yeoh (2020): *Explainable Planning Using Answer Set Programming*. In: *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning, KR 2020, Rhodes, Greece, September 12-18, 2020*, pp. 662–666, doi:10.24963/kr.2020/66.

[3] Tran Cao Son, Van Nguyen, Stylianos Loukas Vasileiou & William Yeoh (2021): *Model Reconciliation in Logic Programs*. In Wolfgang Faber, Gerhard Friedrich, Martin Gebser & Michael Morak, editors: *JELIA*, *Lecture Notes in Computer Science* 12678, Springer, pp. 393–406, doi:10.1007/978-3-030-75775-5_26.

[4] Sarath Sreedharan, Tathagata Chakraborti & Subbarao Kambhampati (2021): *Foundations of explanations as model reconciliation*. *Artif. Intell.* 301, p. 103558, doi:10.1016/j.artint.2021.103558.

[5] Sarath Sreedharan, Alberto Olmo Hernandez, Aditya Prasad Mishra & Subbarao Kambhampati (2019): *Model-Free Model Reconciliation*. In Sarit Kraus, editor: *IJCAI*, ijcai.org, pp. 587–594, doi:10.24963/ijcai.2019/83.

[6] Stylianos Loukas Vasileiou, Alessandro Previti & William Yeoh (2021): *On Exploiting Hitting Sets for Model Reconciliation*. In: *AAAI*. Available at `https://ojs.aaai.org/index.php/AAAI/article/view/16807`.