Generating Explanations from Abstract Argumentation Frameworks: A Goal Selection Case Study (Extended Abstract)

Mariela Morveli-Espinoza$^{[0000-0002-7376-2271]}$ and
Cesar Augusto Tacla$^{[0000-0002-8244-8970]}$

Program in Electrical and Computer Engineering (CPGEI),
Federal University of Technology - Paraná (UTFPR), Curitiba - Brazil
morveli.espinoza@gmail.com, tacla@utfpr.edu.br

Abstract. Generating explanations from logic-based argumentation formalisms is based on reasoning chains that allow to deduce conclusions, which allows to construct structured arguments. In abstract argumentation frameworks (AAFs), neither arguments nor attacks have a defined structure, so there is not an structured way to generate explanations. Thus, this article – which is an extended abstract of [6] – tackles this problem considering as case study the goal selection process in intelligent agents.

Keywords: Explainability · Abstract argumentation · Goal selection.

1 Introduction

In structured argumentation approaches, arguments are constructed using a formal language, which represent the knowledge contained in such arguments. Let us recall that – in most of the approaches – structured arguments consist of a support (set of premisses) and a conclusion such that the conclusion is the result of applying inference steps over the premisses of the support. Since we already have a reasoning chain for supporting a given conclusion c, an explanation (or explanations) for that conclusion can be directly and naturally constructed from the arguments whose conclusions are c. Some explainable approaches based on structured argumentation are the following: [1],[5],[8],[9], and [10]. On the other hand, in abstract argumentation approaches arguments are atomic, that is they do not have an internal structure, which makes difficult the generation of explanations. To the best of our knowledge, there are few articles that tackle this problem (e.g., [2]). In [6], we use an argumentation-based approach for generating explanations about that goal reasoning path in intelligent agents. In this article, we present a case study based on the definitions given in [6]. Although it is a particular case, it can be a first step for a generalized approach since the idea – for future research – is to propose a general approach that can be used in any domain.

In beliefs-desires-intentions (BDI) agents [7], goal selection is a phase of practical reasoning that aims to decide what state of affairs an agent wants to achieve. The input is a set of pursuable goals (or desires) and the output is a set of pursued goals (or intentions) the agent commits to. In [4], an argumentation-based proposal for dealing with goal selection is presented. One of the results is an AAF where arguments represent pursuable goals and attacks the conflicts that exist between pursuable goals, which can be terminal (denoted by t), due to resources (denoted by r), and superfluity (denoted by s). Besides, a semantics was proposed in order to obtain the set of pursued goals. Considering this background, in this article, we aim to generate an explanation for why a goal came (or not) pursued including the form of conflicts (or incompatibility) between goals. This process will be done by constructing structured arguments taking into account structure of the AAF. Next section presents the case study and Section 3 is devoted to some conclusions and the future work.

2 Case Study: Goal Selection

For our case study, we will use the well-known "cleaner world" scenario, where a set of robots (intelligent agents) has the task of cleaning a dirty environment. The main goal of all the robots is to have the environment clean. Besides cleaning, the robots may have other goals such as recharging their batteries or being fixed. Suppose that at a given moment one of the robots (let us call him BOB) detects dirt in slot (5,5); hence, the goal "cleaning (5,5)" becomes pursuable. On the other hand, BOB also auto-detects a technical defect; hence, the goal "be fixed" also becomes pursuable. Suppose that BOB cannot commit to both goals at the same time because the plans adopted for each goal lead to an inconsistency. This means that BOB has to decide because only one of the goals will become pursued. It is natural to think that he can be asked for an explanation about his decision. So, it is important to endow the agents with the ability of explaining their decisions, that is, to explain how and why a certain pursuable goal became (or not) a pursued goal.

In order to generate the explanations, the input will be a Goal AF and the set of pursued goals. Let us consider the following Goal AF: $\mathcal{GAF}_{sc} = \langle \mathcal{G}, \mathcal{RG}_{sc}, \texttt{INCOMP\_G}, \texttt{PREF} \rangle$ where $\mathcal{G}$ is the set of pursuable goals, PREF is the preference relation between goals, $\mathcal{RG}_{sc}$ is the attack relation after considering the preference relation, and $\texttt{INCOMP\_G} : \mathcal{RG}_{sc} \rightarrow 2^{t,r,s}$. Figure 1 shows the $\mathcal{GAF}_{sc}$. Besides, consider that after applying the semantics, the set of pursued goals is $\mathcal{G}' = \{clean(5,5), mop(5,5), be(fixed)\}$.

Firstly, we map the goals in $\mathcal{G}$ into constants of $\mathcal{L}$ in the following manner: $g_1 = clean(5,5)$, $g_2 = pickup(5,5)$, $g_3 = mop(5,5)$, $g_4 = be(in\_workshop)$, and $g_5 = be(fixed)$. We will also map the beliefs and rules to constants in $\mathcal{L}$[1].

[1] The set of rules necessary for the construction of explanations can be found in [6].

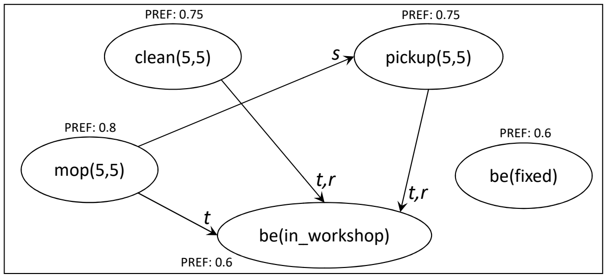


Fig. 1. Goal AF for the cleaner world scenario. The text next to each arrow indicates the form of incompatibility.

We can now follow the steps to generate the explanations:

1. Generate beliefs

- $b_1 : \neg incomp(g_5)$
- $b_2 : incompat(g_3, g_2, `s')$
- $b_3 : incompat(g_3, g_4, `t')$
- $b_4 : incompat(g_1, g_4, `t, r')$
- $b_5 : incompat(g_2, g_4, `t, r')$
- $b_6 : max\_util(g_1)$
- $b_7 : max\_util(g_3)$
- $b_8 : max\_util(g_5)$
- $b_9 : \neg max\_util(g_2)$

$b_{10} : \neg max\_util(g_4)$
$b_{11} : pref(g_3, g_4)$
$b_{12} : \neg pref(g_4, g_3)$
$b_{13} : pref(g_1, g_4)$
$b_{14} : \neg pref(g_4, g_1)$
$b_{15} : pref(g_2, g_4)$
$b_{16} : \neg pref(g_4, g_2)$
$b_{17} : pref(g_3, g_2)$
$b_{18} : \neg pref(g_2, g_3)$

2. Trigger rules

- $r_1 : \neg incomp(g_5) \rightarrow pursued(g_5)$
- $r_2 : incompat(g_3, g_2, `s') \wedge pref(g_3, g_2) \rightarrow pursued(g_3)$
- $r_3 : incompat(g_3, g_2, `s') \wedge \neg pref(g_2, g_3) \rightarrow \neg pursued(g_2)$
- $r_4 : incompat(g_3, g_4, `t') \wedge pref(g_3, g_4) \rightarrow pursued(g_3)$
- $r_5 : incompat(g_3, g_4, `t') \wedge \neg pref(g_4, g_3) \rightarrow \neg pursued(g_4)$
- $r_6 : incompat(g_1, g_4, `t, r') \wedge pref(g_1, g_4) \rightarrow pursued(g_1)$
- $r_7 : incompat(g_1, g_4, `t, r') \wedge \neg pref(g_4, g_1) \rightarrow \neg pursued(g_4)$
- $r_8 : incompat(g_2, g_4, `t, r') \wedge pref(g_2, g_4) \rightarrow pursued(g_2)$
- $r_9 : incompat(g_2, g_4, `t, r') \wedge \neg pref(g_4, g_2) \rightarrow \neg pursued(g_4)$
- $r_{10} : max\_util(g_1) \rightarrow pursued(g_1)$
- $r_{11} : max\_util(g_3) \rightarrow pursued(g_3)$ - $r_{12} : max\_util(g_5) \rightarrow pursued(g_5)$
- $r_{13} : \neg max\_util(g_2) \rightarrow \neg pursued(g_2)$ - $r_{14} : \neg max\_util(g_4) \rightarrow \neg pursued(g_4)$

3. Construct explanatory arguments

- $A_1 = \langle \{b_1, r_1\}, pursued(g_5)\}\rangle$ - $A_2 = \langle \{b_2, b_{17}, r_2\}, pursued(g_3)\}\rangle$
- $A_3 = \langle \{b_2, b_{18}, r_3\}, \neg pursued(g_2)\}\rangle$ - $A_4 = \langle \{b_3, b_{11}, r_4\}, pursued(g_3)\}\rangle$
- $A_5 = \langle \{b_3, b_{12}, r_5\}, \neg pursued(g_4)\}\rangle$ - $A_6 = \langle \{b_4, b_{13}, r_6\}, pursued(g_1)\}\rangle$
- $A_7 = \langle \{b_4, b_{14}, r_7\}, \neg pursued(g_4)\}\rangle$ - $A_8 = \langle \{b_5, b_{15}, r_8\}, pursued(g_2)\}\rangle$
- $A_9 = \langle \{b_5, b_{16}, r_9\}, \neg pursued(g_4)\}\rangle$ - $A_{10} = \langle \{b_6, r_{10}\}, pursued(g_1)\}\rangle$

- $A_{11} = \langle\{b_7, r_{11}\}, pursued(g_3)\}\ \rangle$ - $A_{12} = \langle\{b_8, r_{12}\}, pursued(g_5)\}\rangle$
- $A_{13} = \langle\{b_9, r_{13}\}, \neg pursued(g_2)\}\rangle$ - $A_{14} = \langle\{b_{10}, r_{14}\}, \neg pursued(g_4)\}\rangle$

4. For each goal, generate an explanatory AF and extension

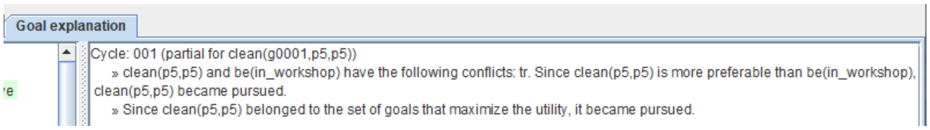
- For g_1: $\mathcal{XAF}_{g_1} = \langle\{A_6, A_{10}\}, \{\}\rangle$, $\mathcal{E} = \{A_6, A_{10}\}$
- For g_2: $\mathcal{XAF}_{g_2} = \langle\{A_3, A_8, A_{13}\}, \{(A_3, A_8), (A_{13}, A_8)\}\rangle$, $\mathcal{E} = \{A_3, A_{13}\}$
- For g_3: $\mathcal{XAF}_{g_3} = \langle\{A_2, A_4, A_{11}\}, \{\}\rangle$, $\mathcal{E} = \{A_2, A_4, A_{11}\}$
- For g_4: $\mathcal{XAF}_{g_4} = \langle\{A_5, A_7, A_9, A_{14}\}, \{\}\rangle$, $\mathcal{E} = \{A_5, A_7, A_9, A_{14}\}$
- For g_5: $\mathcal{XAF}_{g_5} = \langle\{A_1, A_{12}\}, \{\}\rangle$, $\mathcal{E} = \{A_1, A_{12}\}$

Thus, the – partial or complete – explanations for justifying the status of each goal were generated. Next, we present the query, set of arguments of the partial explanation, and the explanatory sentences for the status of goals g_2 and g_3:

- For the query WHY_NOT(g_2), we have $\mathcal{PE} = \{A_3, A_{13}\}$, which can be written:
  * *$\underline{mop(5,5)}$ and $\underline{pickup(5,5)}$ have the following conflicts: ‘$\underline{s}$’. Since $\underline{pickup(5,5)}$ is less preferable than $\underline{mop(5,5)}$, $\underline{pickup(5,5)}$ did not become pursued*
  * *Since $\underline{pickup(5,5)}$ did not belong to the set of goals that maximizes the utility, it did not become pursued*
- For the query WHY(g_3), we have $\mathcal{PE} = \{A_2, A_4, A_{11}\}$, which can be written:
  * *$\underline{mop(5,5)}$ and $\underline{pickup(5,5)}$ have the following conflicts: ‘$\underline{s}$’. Since $\underline{mop(5,5)}$ is more preferable than $\underline{pickup(5,5)}$, $\underline{mop(5,5)}$ became pursued*
  * *$\underline{mop(5,5)}$ and $\underline{be(in\_workshop)}$ have the following conflicts: ‘$\underline{t}$’. Since $\underline{mop(5,5)}$ is more preferable than $\underline{be(in\_workshop)}$, $\underline{mop(5,5)}$ became pursued*
  * *Since $\underline{mop(5,5)}$ belonged to the set of goals that maximizes the utility,* it *became pursued*

For all the queries, except WHY_NOT(g_2), the complete explanation is the same. In the case of WHY_NOT(g_2), the complete explanation includes the attack relations between some of the arguments of its explanatory AF.

We are also working in a simulator for generating explanations [3][2]. In its first version, just partial explanations are generated. Figure 2 shows the explanation for query WHY(g_1).



```
Cycle: 001 (partial for clean(g0001,p5,p5))
    » clean(p5,p5) and be(in_workshop) have the following conflicts: tr. Since clean(p5,p5) is more preferable than be(in_workshop),
clean(p5,p5) became pursued.
    » Since clean(p5,p5) belonged to the set of goals that maximize the utility, it became pursued.
```

Fig. 2. Partial explanation for query WHY(g_1). Obtained by using the simulator ArgAgent.

[2] Available at: https://github.com/henriquermonteiro/BBGP-Agent-Simulator/

3 Final Remarks

This article presents an ongoing work for generating explanations from AAFs. We use the goal selection in intelligent agents as case study. We can notice that atomic arguments can represent more than just reasons, in the case study they represented goals. We aim to study and propose a general model for generating explanations from AAFs considering not only attack relation but other types of relations and concepts like support and preference.

Acknowledgements

The first author is supported by CAPES.

References

1. Cyras, K., Delaney, B., Prociuk, D., Toni, F., Chapman, M., Dominguez, J., Curcin, V.: Argumentation for explainable reasoning with conflicting medical recommendations (2018)
2. Čyras, K., Letsios, D., Misener, R., Toni, F.: Argumentation for explainable scheduling. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 2752–2759 (2019)
3. Jasinski, H., Morveli-Espinoza, M., Tacla, C.: Generating pseudo-natural explanations for goal selection. In: To be published in the Proceedings of the 20th Workshop on Computational Models of Natural Argument (CMNA) (2020)
4. Morveli-Espinoza, M.M., Nieves, J.C., Possebom, A.T., Puyol-Gruart, J., Tacla, C.A.: An argumentation-based approach for identifying and dealing with incompatibilities among procedural goals. International Journal of Approximate Reasoning **105**, 1–26 (2019). https://doi.org/10.1016/j.ijar.2018.10.015, https://doi.org/10.1016/j.ijar.2018.10.015
5. Morveli-Espinoza, M., Possebom, A., Tacla, C.A.: Argumentation-based agents that explain their decisions. In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS). pp. 467–472. IEEE (2019)
6. Morveli-Espinoza, M., Tacla, C.A., Henrique, H.: An argumentation-based approach for explaining goals selection in intelligent agents. In: 2020 9th Brazilian Conference on Intelligent Systems (BRACIS) (2020)
7. Rao, A.S., Georgeff, M.P.: BDI agents: from theory to practice. In: ICMAS. vol. 95, pp. 312–319 (1995)
8. Sassoon, I., Sklar, E., Kokciyan, N., Parsons, S.: Explainable argumentation for wellness consultation. In: Proceedings of 1st International Workshop on eXplanable TRansparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS2019), AAMAS (2019)
9. Zeng, Z., Fan, X., Miao, C., Leung, C., Jih, C.J., Soon, O.Y.: Context-based and explainable decision making with argumentation. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. pp. 1114–1122. International Foundation for Autonomous Agents and Multiagent Systems (2018)
10. Zhong, Q., Fan, X., Luo, X., Toni, F.: An explainable multi-attribute decision model based on argumentation. Expert Systems with Applications **117**, 42–61 (2019)