Dynamic Argumentation and Action Languages: Towards Explanations

Yann Munro, Camilo Sarmiento, Isabelle Bloch, Gauvain Bourgne, Marie-Jeanne Lesot

Dynamic Argumentation and Action Languages: Towards Explanations

Yann Munro, Camilo Sarmiento,
Isabelle Bloch, Gauvain Bourgne, and Marie-Jeanne Lesot

Sorbonne Université, CNRS, LIP6, Paris, France
firstname.surname@lip6.fr

Abstract. In the context of abstract argumentation, we present the benefits of considering temporality, i.e. the order in which arguments are enunciated, so as to better model dialogues, reason about causal relationships between variables and provide explanations. Based on a formal method to rewrite the concepts of acyclic abstract argumentation frameworks in an action language, we discuss new ways to generate explanations based on a dynamic dialogue. As this modelling offers tools for visually narrating the evolution of the dialogue and the acceptability of its arguments, it is a good starting point for studying human-centred explanations. In addition, we investigate how richer causal relations can be used to provide explanations in addition to the graphical representation. Yet we observe that the extracted causal chains can be very long and contain redundancies, and thus do not necessarily provide satisfactory explanations. We therefore discuss directions to derive human-centred explanations from these causal chains satisfying recommendations provided by cognitive science studies.

Keywords: Argumentation · Action Languages · Causality · XAI.

1 Introduction

The abstract argumentation framework (AAF), first introduced by Dung [5], provides a suitable framework for representing and reasoning about contradicting information using arguments and a binary relation between them called the attack relationship. It makes it possible to find sets of arguments, called extensions, that can be accepted together, and then to give explanations on why such sets have been accepted or not. Now, if one wishes to model dialogues and provide both on-the-spot explanations and human-centred explanations, it is essential to be able to keep track of the order in which arguments have been stated, and thus to include a notion of temporality. This is not possible with the classic version of AAF as it is a static framework, unable to model and reason about time.

Several types of dynamic extended argumentation frameworks have been proposed to tackle these issues. The YALLA logical formalism [12] proposes to rewrite an AAF and then use revision or belief change operators to update the argumentation system. It offers a very expressive language that allows finding which attack relations or arguments should be removed or added in order to achieve a given goal in the subsequent time step. Such a reasoning can be used to build an explanation. However, this approach

seems not very suited to our problem given that we are aiming to model all the dialogue, and in YALLA the number of terms increases exponentially with the number of arguments. Another approach by Doutre et al. [4] proposes to use a dynamical propositional language to model the argumentation graph and its evolution across the dialogue. By doing so, it partially answers our question by successfully integrating the notion of temporality. Since the aim of this long-term work is to propose an explanation of why an argument is or is not acceptable at a given moment in the dialogue, this proposal is not sufficient. Indeed, to achieve that, a formal study of the notion of causality is required.

Even if there is no single definition of an explanation, it is accepted from a social science point of view that an explanation is an answer to a *why* question as reported by Miller in [8]. Therefore, it is deeply related to the notion of causality, justifying why a formal study of this notion is required. Causality has already been studied for AAFs. Recently, Bengal et al. [2] used a knowledge-based formalism to define counterfactual reasoning for structured argumentation. Their definition is the direct application of Pearl's definition for causal model [11], similar to what is called a *but-for* test. Before that, Bochman in [3] established an equivalence between abstract argumentation and a causal reasoning system. This system has latter be extended for more complex causal reasoning. However, all these works are about classical AAFs without temporality.

This paper proposes to investigate the use of another logical formalism, namely action languages [6,7], that have been created for modelling temporality and include a formal study of causality [14], both for modelling the dynamics of a dialogue and for generating explanations. Action languages offer tools to reason about action and change and have been naturally conceived to include the notion of time. The action language introduced by Sarmiento et al. [14] has been designed to determine the evolution of the world given a set of actions corresponding to deliberate choices of the agent, the occurrence of which can trigger a chain reaction through external events. It is a classical transition system: the set of events which occur at time point t, denoted $E(t)$, generates the transition between the states $S(t)$ and $S(t+1)$. Thus, the states follow one another as events occur, simulating the evolution of the world. In order to have a complete knowledge of the evolution of the world, two traces are generated: one tracing the evolution of world states, the other tracing the occurrence of events.

Furthermore, this action language includes tools to reason about causality. A causal relation links a cause to an effect. As commonly accepted by philosophers, in an actual causality relation, both the cause and the effect are occurrences of events [1]. Since action languages represent the evolution of the world as a succession of states produced by the occurrence of events, states are introduced between events. Therefore, it is necessary to define causal relations where causes are occurrences of events and effects are properties of the world being true at a specific time point. The actual causation definition proposed by Sarmiento et al. [14] is an action language suitable formalisation where these intermediate relations are established on the basis of Wright's NESS test of causation [15]. A sound and complete translation into ASP has been proposed [13]. We propose to take advantage of these properties, in particular the capacity to reason about causality, to study the causal relations in a dialogue represented as a dynamic graph, paving the way for the search of explanations.

This work constitutes a first step of a proposition to bring together different existing tools from the Knowledge Representation and Reasoning field to tackle issues that arise in abstract argumentation when considering causality and temporality together. The aim of this paper being to discuss the possibilities offered by the cross-fertilisation of action languages, causality and abstract argumentation, technical details are omitted. For interested readers, an available version [10] details the formalisation of acyclic abstract argumentation graphs into an action language and some of its formal properties, including its soundness and completeness, as well as the relevance of the temporality inclusion. Section 2 briefly recalls the elements proposed in [10]. Section 3 presents the main contributions of this paper: exploring how the use of action languages can benefit AAFs to generate explanations.

2 From Argumentation to Action Language

In order to transform an argumentation framework into an action language, the first step we propose in [10] is to define *fluents*, i.e. the variables which describe the state of the world. In this paper, the world consists only of arguments and the relation between them. Thus, only few fluents are necessary to describe Dung's AAF completely: p_x, a_x and $cA_{y,x}$, respectively representing that argument x is present in the dialogue i.e has been enunciated, x is acceptable, and argument y attacks x.

As the objective is to inject time and thus model dialogues better, the second step is to define the *events* that will create change in the world. During a debate, the only possible deliberate *action* is enunciating an argument x, denoted $enunciate_x$. However, even if this is the only possible intentional action, enunciating an argument can lead to changes in the acceptability of other arguments. This phenomenon is modelled using so-called *exogenous events*, events that are triggered as soon as some specific conditions are verified without the need of an agent to perform them. The first one, called $makesUnacc_{x,y}$, represents the fact that argument x is directly making y unacceptable. The second one, $makesAcc_y$, models the case where y becomes acceptable again. This can happen because an argument x just made an argument attacking y not acceptable, and so y is not attacked by any acceptable arguments anymore, leading to the acceptability of y.

Munro et al. established [10] that this transformation is sound and complete in the case of an acyclic graph. It means that, when the dialogue is over and the action language has finished all the acceptability updates, an argument x that is acceptable in the corresponding argumentation graph has its corresponding fluent a_x being true and vice-versa. This property proves that the formalisation leads neither to a loss of information nor creating new one.

Furthermore, the addition of a notion of time in the framework enriches the model. This is illustrated and proven in [10] using the traces: given a set of arguments, a set of attacks, and a sequence, the obtained state and event traces are unique, meaning that changing the order of enunciation leads to a different story line but still the same end. Moreover, as the time line changes, even if the end is still unique, the events leading to this end are also different, and thus its causes as well. That is the reason why it is essential to take temporality into account when dealing with notions close to causality,

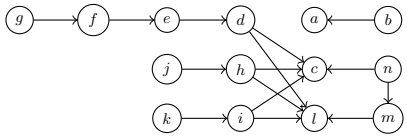

Fig. 1. Argumentation graph associated with Example 1.

especially in explainability. An ASP implementation of those theoretical tools is available[1] and can be used as a basis to generate graphical representations such as those shown in the next section.

3 Towards Explanations: a Three Level Process

By formalising acyclic abstract argumentation graphs into an action language as briefly recalled in Section 2, we now have a complete knowledge of the evolution of the dialogue. Indeed, we are able to generate two traces: one tracing the evolution of world states, the other tracing the occurrence of events. This section proposes to explore at three levels how the use of action languages can benefit AAFs to generate explanations. Example 1 is used throughout this section to illustrate the discussion. The state trace corresponding to the example counts thirty one states and the event trace counts fourteen actions and twenty exogenous events.

Example 1. To illustrate these notions, Munro et al. [10] introduced an argumentative scenario modelling the interaction between a requesting physician, D, and a radiologist, R, concerning an examination of a baby for a pathology Z. After an initial discussion between the two specialists, which we will not go into here, the decision was taken to schedule an MRI scan in two days' time. But later that day, the physician receives a call from the family saying that the baby is really not well and insisting on the urgency of the examination. Therefore, the doctor contacts the radiologist to add a final argument. From the detailed dialogue, we can extract manually arguments and their relations to create the AAF represented in Figure 1 with the following arguments: {**a**: Scanner, b: Ionising radiation, **c**: MRI in two days, d: Z not visible by MRI, e: Z visible by MRI, f: Difficult conditions, g: High experience, h: High cost for the hospital, i: High cost for the patient, j: Not problematic for the hospital, k: The family is covered for an MRI, **l**: MRI today, m: No availability today, n: It is an emergency!}. Arguments a, c, l are called the decision variables, their acceptance being the criterion leading to a decision.

3.1 Temporal Modelling and Graphical Representation

Transition systems such as action languages offer a way to model the world that is suitable for visual representation. Indeed, the representation of two successive states and the events that led to this evolution gives an accurate representation of how the evolution of the world is modelled. For that, in the case of the considered action language, we use

[1] https://gitlab.lip6.fr/sarmiento/temporality_and_causality_in_abstract_argumentation.git

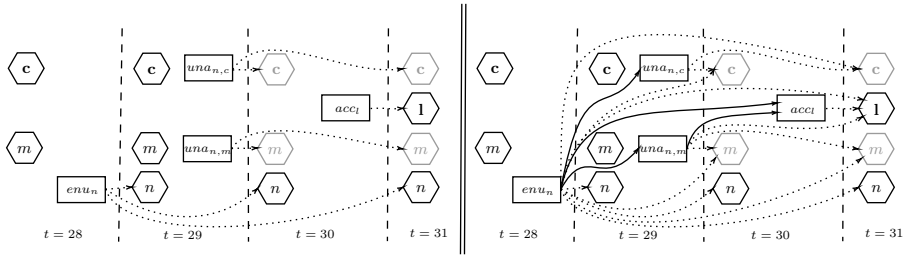

Fig. 2. (Left) Partial and (right) enriched graphical representations of an illustrative example of event and state traces, with extracted causal relations: NESS-causes (· · ·) and actual causes (—).

the traces. Indeed, they record the evolution of the world. Therefore, displaying them in chronological order provides a narration of the interaction.

Following this idea of using the narration of the interaction provided by traces, in Figure 2, we propose a possible display of the event and state traces obtained from Example 1 by showing fluents as hexagons and triggered events as rectangles. Since the acceptability of arguments is what mainly matters in a dialogue, we represent only the fluents a_x, using only the argument names for the sake of legibility. Moreover, we do not show fluents when their negation holds in the state, except when the occurrence of a represented event results in the negation of the fluent. In this case, the negation is represented by a lighter shade. The events $enunciate_x$, $makesUnacc_{y,x}$, and $makesAcc_x$ are shortened as enu_x, $una_{y,x}$, and acc_x, respectively.

The partial graphical representation (left of Figure 2) shows the causal relations that can be deduced immediately, these are the direct causal relations. They can be used to derive directly basic explanations such as the fact that enunciating argument n at time $t = 28$ causes n to be acceptable in the next state, i.e. at time $t = 29$. Or, if one wonders why argument l has become acceptable at $t = 31$, such a representation shows that it is because $makesAcceptable_l$ triggered. In contrast to the case of n, which we find acceptable due to its recent enunciation, the present causal relationship indicates l's acceptability has now been restored, which means that all of its attackers must have been rendered unacceptable. Thus, using the temporal modelling that is offered by the action language permits to give a visual display of the dialogue, helping to increase the understanding of the interaction.

However, using only the temporality to derive simple causal relations between states is not enough and more complex relations are needed. Indeed, on the acceptability of argument l, the basic causal relation linking it with $makesAcceptable_l$ is not satisfactory enough. To solve this issue, causal reasoning is needed to find more complex and relevant causes, called actual causes.

3.2 Causal Reasoning

In order to give an actual causality definition suitable for action languages, two types of causal relations are introduced in [14]: (i) *NESS-causes* relate occurrences of events and

properties of the world being true at a specific time point, i.e fluents; (ii) *Actual causes* relate two occurrences of events. In particular, the occurrence of event e is considered an actual cause of the occurrence of event e' if and only if the occurrence of e is a NESS-cause of the conditions necessary to the triggering of e'. The transitivity of the actual cause relation leads to build a causal chain that allows us to go back in time to track all causes.

We propose, on the right part of Figure 2, a richer graphical representation to visualise these enhanced causal relations. In this one, more complex relationships appear, relationships that can be used for explanation. For example, the richer causal relations tell us that the enunciation of n is a cause of l being acceptable and c and m not being acceptable, i.e. the fact that it is an emergency causes doing the MRI today instead of in three days despite there is officially no available slot left. Moreover, by looking at the actual causal relations, we can understand, for example, that the enunciation of n causes l to be acceptable by making m not acceptable, which is much more satisfactory. More complex relationships of this kind can then be obtained by reconstructing the chain backwards. This can help to explain the dialogue in a simple way when the number of arguments and attack relations between them becomes large.

3.3 Towards Explanations

These graphical representations, like the causal chains, may not represent an explanation per se. Indeed, these relations, as helpful as they can be, tend to form long causal chains including redundancy or useless variables. This raises the question of the appropriate way to use these causal chains to derive explanations. The aim of the current work is therefore to propose a method that makes it possible to extract, from this enrichment of argument graphs, explanations on the acceptability or non-acceptability of one or more arguments. A first direction can be found in the works that study the links between the notions of causality and explanation. These are summarised in Miller's article [8], which establishes several interesting properties that an explanation should satisfy. We will discuss five of them: its proximity to the consequence, its consideration of an agent's volition, its contrastivity, its robustness, and its short length. The question is then how to apply these principles to the new framework we propose in order to provide explanations to a human user.

Some of these principles seem simple to apply in the action language. First, its temporal proximity to the consequence is easy to evaluate and formalise thanks to the inclusion of time in the framework. Another interesting and easy to include property is that a deliberate action is often preferred to an exogenous event. Then, given two identical explanations except for one element, the one with the temporal closest deliberate actions will be preferred.

The next two properties, contrastivity and robustness, are more complex to integrate. First of all, following what Miller did with Halpern's structural causal model [9], to define a contrastive explanation, we first need to introduce the formalism for a contrastive causal chain. Intuitively, contrastive causes of a_l and $\neg a_m$ would be the common elements of the two causal chains. Then, regarding the robustness, an explanation can be considered more robust than another if it holds in a larger number of scenarios. Finally, an explanation has to be short.

Following those principles, an explanation for doing an IRM today instead of in two days could be the enunciation of n stating that it is an emergency. Indeed, it is a short and contrastive explanation made of a deliberate action, temporally close to the consequence. In this case, there is an explanation satisfying all the desirable properties. Otherwise, to provide a "good" explanation, the question of how to aggregate these different properties will have to be addressed.

4 Directions for Future Work

As this work is a first step in exploring how the use of action languages can benefit AAFs to generate explanations, some simplifying assumptions have been made. Future works will focus on generalising the transformation into the action language framework to richer argumentation cases, in particular cyclic argumentative graphs. Ongoing works also aim at formalising Miller's desired explanation properties in the action language framework so as to propose a compliant explanation generation and ordering method. The next step will aim at evaluating this method experimentally, conducting user studies to assess the intelligibility of the generated explanations, both in terms of objective understanding and subjective satisfaction.

References

1. Andreas, H., Guenther, M.: Regularity and Inferential Theories of Causation. In: The Stanford Encyclopedia of Philosophy. Stanford University (2021)
2. Bengel, L., Blümel, L., Rienstra, T., Thimm, M.: Argumentation-based causal and counterfactual reasoning. In: 1st International Workshop on Argumentation for eXplainable AI (2022)
3. Bochman, A.: Propositional argumentation and causal reasoning. In: International Joint Conference on Artificial Intelligence. vol. 19, p. 388 (2005)
4. Doutre, S., Maffre, F., McBurney, P.: A dynamic logic framework for abstract argumentation: adding and removing arguments. In: Advances in Artificial Intelligence: From Theory to Practice: 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017. Springer (2017)
5. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**, 321–357 (1995)
6. Fox, M., Long, D.: Modelling Mixed Discrete-Continuous Domains for Planning. Journal of Artificial Intelligence Research **27**, 235–297 (2006)
7. Giunchiglia, E., Lifschitz, V.: An Action Language Based on Causal Explanation: Preliminary Report. In: 15th Nat. Conf. on Artificial Intelligence and 10th Innovative Applications of Artificial Intelligence Conf., AAAI 98, IAAI 98. pp. 623–630 (1998)
8. Miller, T.: Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence **267**, 1–38 (2019)
9. Miller, T.: Contrastive explanation: a structural-model approach. The Knowledge Engineering Review **36** (2021)
10. Munro, Y., Sarmiento, C., Bloch, I., Bourgne, G., Lesot, M.J.: Temporality and causality in abstract argumentation. arXiv, CoRR **abs/2303.09197** (2023)
11. Pearl, J., et al.: Models, reasoning and inference. Cambridge, UK: Cambridge University Press **19**(2), 3 (2000)

12. Dupin de Saint-Cyr, F., Bisquert, P., Cayrol, C., Lagasquie-Schiex, M.C.: Argumentation update in YALLA (yet another logic language for argumentation). International Journal of Approximate Reasoning **75**, 57–92 (2016)
13. Sarmiento, C., Bourgne, G., Inoue, K., Cavalli, D., Ganascia, J.G.: Action Languages Based Actual Causality for Computational Ethics: a Sound and Complete Implementation in ASP. arXiv, CoRR **abs/2205.02919** (2023)
14. Sarmiento, C., Bourgne, G., Inoue, K., Ganascia, J.G.: Action Languages Based Actual Causality in Decision Making Contexts. In: PRIMA. vol. LNCS 13753. Springer (2022)
15. Wright, R.W.: Causation in Tort Law. California Law Review **73**(6), 1735–1828 (1985)