Contrastive Explanations for Argumentation-Based Conclusions

Extended Abstract

AnneMarie Borg
Utrecht University
Utrecht, The Netherlands
a.borg@uu.nl

Floris Bex
Utrecht University
Utrecht, The Netherlands
Tilburg University
Tilburg, The Netherlands
f.j.bex@uu.nl

ABSTRACT

In this paper we discuss *contrastive explanations* for formal argumentation – the question why one argument (the fact) can be accepted, whilst another argument (the foil) cannot be accepted. We show under which conditions contrastive explanations in abstract argumentation are meaningful, and how argumentation allows us to make implicit foils explicit.

KEYWORDS

Formal Argumentation; Explainable Artificial Intelligence

ACM Reference Format:
AnneMarie Borg and Floris Bex. 2022. Contrastive Explanations for Argumentation-Based Conclusions: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

Explainable AI (XAI) has become an important research direction in AI [10]. AI systems, including formal argumentation [1], are being applied in a variety of real-life situations and therefore require an explanation method. A number of methods for determining explanations for the (non-)acceptability of arguments have been proposed [4]. What is still lacking, however, is an argumentation-based interpretation of *contrastive explanations*.

Contrastiveness is central to explanations [6–8]: when people ask *'Why P?'*, they often mean *'Why P rather than Q?'* – here P is called the *fact* and Q is called the *foil* [6]. The answer to the question is then to explain as many of the differences between fact and foil as possible. However, in formal argumentation the existing work focuses on *'Why is argument A (not) acceptable?'* instead of the contrastive question *'Why is argument A acceptable and argument B not?'* (or vice versa) and no work on contrastiveness exists.

In this paper we extend the basic framework from [2] with which explanations for accepted and non-accepted arguments or formulas can be formulated in a variety of ways. The introduced contrastive explanations return the common elements of the acceptance explanation of the fact and the non-acceptance explanation of the foil. We show that in almost all situations these explanations are meaningful, i.e., that such common elements exist. Additionally we show that we can provide contrastive explanations when the foil is not explicitly known.[1]

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

2 PRELIMINARIES

We focus on explanations for conclusions derived from Dung-style argumentation frameworks.

An *abstract argumentation framework* (AF) [5] is a pair $\mathcal{AF} = \langle \text{Args}, \text{Att} \rangle$, where Args is a set of *arguments* and $\text{Att} \subseteq \text{Args} \times \text{Args}$ is an *attack relation* on these arguments. An argumentation framework can be viewed as a directed graph, in which the nodes represent arguments and the arrows represent the attacks.

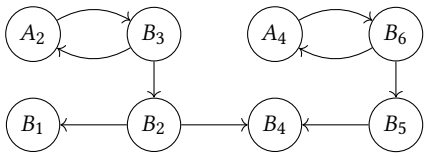

Figure 1: Graphical representation of the AF $\mathcal{AF}_1$.

Example 2.1. Figure 1 represents the argumentation framework $\mathcal{AF}_1 = \langle \text{Args}_1, \text{Att}_1 \rangle$ where $\text{Args}_1 = \{A_2, A_4, B_1, B_2, B_3, B_4, B_5, B_6\}$ and $\text{Att}_1 = \{(A_2, B_3), (A_4, B_6), (B_2, B_1), (B_2, B_4), (B_3, A_2), (B_3, B_2), (B_5, B_4), (B_6, A_4), (B_6, B_5)\}$.

Dung-style semantics [5] can be applied to an AF, to determine the sets of arguments (called *extensions*) that can be accepted. In this abstract we will work with preferred semantics. An argument is *accepted* if it is part of a preferred extension and it is *not accepted* if there is a preferred extension without that argument.

We will require that the explanation for an argument A is *relevant* (i.e., the arguments in the explanation (in)directly attack or defend A), in order to prevent that explanations contain arguments that do not influence the acceptance of A. We will say that an argument B is *conflict-relevant* for A if B (in)directly attacks A.

In what follows we assume that we have an acceptance and a non-acceptance explanation for arguments, as introduced in [2]. In particular, given an argument A, the acceptance explanation (denoted by $\text{Acc}(A)$) collects the arguments from an extension that defend A against some attack and the non-acceptance explanation (denoted by $\text{NotAcc}(A)$) collects the arguments that attack A and to which an extension does not provide a defense.

Example 2.2. For $\mathcal{AF}_1$ we have that $\text{Acc}(B_2) = \{A_2\}$, $\text{Acc}(B_4) = \{B_3, B_6\}$, $\text{NotAcc}(B_2) = \{B_3\}$ and $\text{NotAcc}(B_4) = \{A_2, A_4, A_2, B_5\}$.

[1] See [3] for the full version of this paper.

3 CONTRASTIVE EXPLANATIONS

A contrastive explanation explains A by explaining *why A rather than B*. Important in contrastive explanations is that the difference between fact (i.e., A) and foil (i.e., B) is highlighted. In this paper we assume that fact and foil are not always compatible: A and B are not always part of the same extension. Intuitively, we make this assumption since otherwise there is no contrastive question for fact and foil (i.e., *why both A and B* is not contrastive).

Contrastive explanations are modeled by comparing the elements of the basic explanations that explain the acceptance of the fact and, at the same time, explain the non-acceptance of the foil.

Definition 3.1 (Contrastive explanations). Let $\mathcal{AF} = \langle \text{Args}, \text{Att} \rangle$ be an AF, let $A \in \text{Args}$ (the fact) and let $\text{S} \subseteq \text{Args}$ (a set of foils) such that there is no preferred extension $\mathcal{E}$ in which $A, B \in \mathcal{E}$ for all $B \in \text{S}$. Contrastive explanations are then defined as $\text{Cont}(A, \text{S}) = \begin{cases} \text{Acc}(A) \cap \bigcup_{B \in \text{S}} \text{NotAcc}(B) & \text{if } \text{Acc}(A) \cap \bigcup_{B \in \text{S}} \text{NotAcc}(B) \neq \emptyset \\ \langle \text{Acc}(A), \bigcup_{B \in \text{S}} \text{NotAcc}(B) \rangle & \text{otherwise.} \end{cases}$

In words, when there are arguments that cause the fact to be accepted and the foil to be non-accepted, the contrastive explanation is the set of such arguments, the first case. If there are no common causes for the acceptance of the fact and the non-acceptance of the foil, the explanation is a pair of the respective explanations, the second case.

Example 3.2. For $\mathcal{AF}_1$ we have the following: $\text{Cont}(B_4, B_2) = \{B_3\}$, $\text{Cont}(B_4, B_5) = \{B_6\}$ and $\text{Cont}(B_4, \{B_2, B_5\}) = \{B_3, B_6\}$.

Recall (Example 2.2) that the acceptance of B_4 can be explained by B_3 and B_6, when compared to the non-acceptance of B_2 [resp. B_5] the acceptance of B_4 is explained by B_3 [resp. B_6] alone.

One could consider these explanations more meaningful when they return a set, rather than a pair. This is the case since then there are arguments that influence both the acceptance of the fact and the non-acceptance of the foil. The next proposition shows that in most cases the explanation is a set. Only when the accepted argument is not attacked or fact and foil are not conflict-relevant is the intersection empty.

Proposition 3.3. *Let $\mathcal{AF} = \langle \text{Args}, \text{Att} \rangle$ be an AF and $A, B \in \text{Args}$. If $\text{Acc}(A) \cap \text{NotAcc}(B) = \emptyset$ then $\text{Acc}(A) = \emptyset$; or A is not conflict-relevant for B.*

In view of the above result, the following conditions are introduced on the fact and foil. By requiring these conditions to hold, meaningful contrastive explanations can be obtained. For this let $\mathcal{AF} = \langle \text{Args}, \text{Att} \rangle$ be an AF and let $\{A\} \cup \text{S} \subseteq \text{Args}$. Then $\text{Cont}(A, \text{S})$ can be requested when, for each $B \in \text{S}$:

- A is at least accepted and B is at least not accepted;
- for each preferred extension $\mathcal{E}$ it never holds that $\{A, B\} \subseteq \mathcal{E}$;
- A is conflict-relevant for B or B is conflict-relevant for A.

These conditions ensure that fact and foil are incompatible, but still relevant for each other: it is explained what makes the fact accepted and, simultaneously causes the foil to be non-accepted. This prevents contrastive explanations for arguments that are not related or conflicting. These conditions are not exhaustive, depending on, e.g., the application, a user might wish to enforce further conditions on fact or foil.

3.1 Non-Explicit Foil

When humans request a (contrastive) explanation the foil is sometimes left implicit, yet the expected explanation does not provide all reasons for the fact happening, but should rather explain the difference between fact and foil. While humans are able to detect the foil based on, e.g., context, this is a challenge for AI systems, including argumentation. In particular, it is impossible to provide one strategy, since different applications entail different foils. For example, if argumentation is applied to determine a yes or no answer for argument A (e.g., whether one qualifies for a loan), the foil would be *not A*, but if the foil should be chosen from a larger set (e.g., a medical diagnosis), it might be any member of that set.

Since in the definition of contrastive explanations it is necessary to provide a foil, a way to determine the foil is required. This is where one of the advantages of formal argumentation comes in: the explicit nature of conflicts between arguments makes that the foil or a set of foils can be constructed from an AF. Since the relation between arguments is only determined by the attack relation in our setting, it is impossible to distinguish between attackers. One example collects all directly attacking arguments.

Definition 3.4. Let $\mathcal{AF} = \langle \text{Args}, \text{Att} \rangle$ be an AF and let $A \in \text{Args}$. Then: $\text{Foil}(A) = \{B \in \text{Args} \mid B \text{ directly attacks } A\}$.

Example 3.5. For the framework $\mathcal{AF}_1$ we have that: $\text{Foil}(B_4) = \{B_2, B_5\}$; $\text{Foil}(B_2) = \{B_3\}$ and $\text{Foil}(B_5) = \{B_6\}$.

In what follows it will be assumed that $\text{Foil}(A) \neq \emptyset$, for fact A, i.e., that a foil exists. Note that, by Definition 3.4, for any AF $\mathcal{AF} = \langle \text{Args}, \text{Att} \rangle$ and $A \in \text{Args}$, $\text{Foil}(A) = \emptyset$ iff there is no $B \in \text{Args}$ such that $(B, A) \in \text{Att}$. Hence, any argument without a foil is not attacked at all. The next proposition shows that the obtained contrastive explanations are meaningful when the first condition of the applicability of contrastive explanations is fulfilled and the foil is defined as in Definition 3.4.

Proposition 3.6. *Let $\mathcal{AF} = \langle \text{Args}, \text{Att} \rangle$ be an AF, let $A \in \text{Args}$ be such that $\text{Foil}(A) \neq \emptyset$. Then a contrastive acceptance explanation can be requested for A, when A is at least accepted and for all $B \in \text{Foil}(A)$, B is at least not accepted.*

In view of the above proposition we obtain the following corollary from Propositions 3.3 and 3.6.

Corollary 3.7. *Let $\mathcal{AF} = \langle \text{Args}, \text{Att} \rangle$ be an AF, let $A \in \text{Args}$ be such that $\text{Foil}(A) \neq \emptyset$. Then: the explanation $\text{Cont}(A, \text{Foil}(A))$ is never of the form $\left\langle \text{Acc}(A), \bigcup_{B \in \text{Foil}(A)} \text{NotAcc}(B) \right\rangle$;*

4 CONCLUSION

In this paper we have introduced a general approach to derive contrastive explanations from AFs generated from an abstract setting. In [3] we consider additional semantics, two additional notions of (non-)acceptance, contrastive explanations for structured settings (i.e., ASPIC$^+$ [9]) as well as a real-life example from an argumentation-based system employed at the Netherlands Police. To the best of our knowledge this is the first investigation into contrastive local explanations for conclusions derived from both abstract and structured argumentation.

Acknowledgements. This research has been partly funded by the Dutch Ministry of Justice and the Netherlands Police.

REFERENCES

[1] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Simari, Matthias Thimm, and Serena Villata. 2017. Towards Artificial Argumentation. *AI magazine* 38, 3 (2017), 25–36.

[2] AnneMarie Borg and Floris Bex. 2021. A Basic Framework for Explanations in Argumentation. *IEEE Intelligent Systems* 36, 2 (2021), 25–35. doi: 10.1109/MIS.2021.3053102.

[3] AnneMarie Borg and Floris Bex. 2021. Contrastive Explanations for Argumentation-Based Conclusions. *CoRR* abs/2107.03265v2 (2021). arXiv:2107.03265v2 https://arxiv.org/abs/2107.03265v2

[4] Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. 2021. Argumentative XAI: A Survey. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI'21)*, Zhi-Hua Zhou (Ed.). ijcai.org, 4392–4399.

[5] Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77, 2 (1995), 321–357.

[6] Peter Lipton. 1990. Contrastive Explanation. *Royal Institute of Philosophy Supplement* 27 (1990), 247–266.

[7] Tim Miller. 2018. Contrastive Explanation: A Structural-Model Approach. *CoRR* abs/1811.03163 (2018). http://arxiv.org/abs/1811.03163

[8] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[9] Henry Prakken. 2010. An abstract framework for argumentation with structured arguments. *Argument & Computation* 1, 2 (2010), 93–124.

[10] Wojciech Samek and Klaus-Robert Müller. 2019. Towards Explainable Artificial Intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (Eds.). Springer, 5–22.