ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-1005

On Computing Explanations in Abstract Argumentation

Xiuyi Fan and **Francesca Toni**[1]

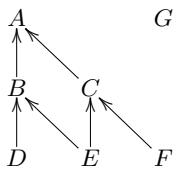
Abstract. Argumentation can be viewed as a process of generating *explanations*. We propose a new argumentation semantics, *related admissibility*, for closely capturing explanations in Abstract Argumentation, and distinguish between *compact* and *verbose* explanations. We show that *dispute forests*, composed of *dispute trees*, can be used to correctly compute these explanations.

1 Introduction

One of the core advantages of argumentation is in transparently explaining the process and results of reasoning. Existing argumentation semantics are designed to answer the question: *Given a set of arguments, which subsets theerof are "good"*? They are less useful in directly answering the question: *Given a set of arguments, why is a particular argument therein "good"*? Of course this question can be answered with "because it belongs to a good set", but this does not provide a direct explanation tailored to the argument in question.

We propose a new argumentation semantics, *related admissibility*, specifically for generating two kinds of explanations in Abstract Argumentation (AA) [1] and present a sound and complete computational counterpart for this semantics using *dispute forests* composed of *dispute trees* as defined in [2]. The following example, adapted from [3], illustrates the motivation behind this work.

Example 1. An agent needs to decide on accommodation in London, amongst three options: Imperial College Student Accommodation (ic), John Howard Hotel (jh), and Ritz Hotel (ritz). The two main criteria for deciding are whether accomodation is cheap and quiet. The agent believes that ic is cheap and quiet, jh is neither, and ritz is only quiet. Also, it believes that London has good public transport. The decision to choose ic can be represented by the following AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$ (as conventional, represented as a directed graph with nodes being arguments in $\mathcal{A}$ and arcs being attacks in $\mathcal{R}$):



A: Choose ic.
B: Why not jh?
D: Because it is not quiet.
C: Why not ritz?
E: But ritz is not cheap (and neither is jh).
F: Also, ritz is fully booked.
G: London has good public transport.

In AA, a set of arguments is *admissible* iff it does not attack itself and it counter-attacks all attacking arguments [1]. Thus, $S_1 = \{A,D,E,F\}$, $S_2 = \{A,D,F\}$, $S_3 = \{A,E\}$ are admissible (whereas, e.g., $\{A\}$ is not). $S_i \cup \{G\}$ is also admissible (for $i = 1, 2, 3$). However, for the purposes of explaining A, G is irrelevant and should not be included. Moreover, S_1 is *verbose*, in that it includes all relevant reasons for A, whereas S_2 and S_3 are *compact*, in that none of the reasons for A they contain can be eliminated from them.

[1] Imperial College London, UK. E-mail: {x.fan09,ft}@imperial.ac.uk

2 Explanations

Giving a general theory for the explanation of human actions and beliefs is a challenging task [4]. It is widely acknowledged that an explanation should be a *justification* [4]:

> *... if I am asked to explain why I hold some general belief that* p*, I answer by giving my justification for the claim that* p *is true.*

Hence, if a belief q does not contribute to the justification of p, q should not be in the explanation of p. This intuition can be given in argumentation terms using a 'defends' relation, defined as follows (here and throughout we assume as given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$):

Definition 1. Let $X, Y \in \mathcal{A}$. X *defends* Y iff:
1. $X = Y$; or
2. $\exists Z \in \mathcal{A}$, s.t. X attacks Z and Z attacks Y; or
3. $\exists Z \in \mathcal{A}$, s.t. X defends Z and Z defends Y.

$S \subseteq \mathcal{A}$ *defends* $X \in \mathcal{A}$ iff $\forall Y \in S$: Y defends X.

Definition 1 is given recursively with (1) and (2) the base cases. Note that each argument defends itself (by (1)).

Example 2. (Example 1 continued.) Each of A, D, E and F defends A, and $\{A, D, E, F\}$ and all its non-empty subsets defend A.

By combining our 'defends' relation and standard admissibility we obtain our notion of related admissible sets of arguments as follows:

Definition 2. A set of arguments $S \subseteq \mathcal{A}$ is *related admissible* iff $\exists X \in S$ s.t. S defends X and S is admissible. Any such X is referred to as a *topic* of S.

Example 3. (Example 1 continued.) $\{A, D, E, F\}$, $\{A, D, E\}$, $\{A, D, F\}$, $\{A, E, F\}$, and $\{A, E\}$ are related admissible, with A the topic of all. $\{F, G\}$ is admissible but not related admissible, since F does not defend G and vice versa.

All arguments in a related admissible set are topics of some related admissible subset thereof. Formally:

Proposition 1. Given a related admissible set $S \subseteq \mathcal{A}$, for all $X \in S$ there is a related admissible set $S' \subseteq S$ s.t. X is a topic of S'.

As an illustration, in Example 3, given $\{A, D, E, F\}$, the related admissible subset whose topic is D is $\{D\}$.

We use the notion of related admissible set to define explanations:

Definition 3. For any argument $X \in \mathcal{A}$, an *explanation* of X is $S \subseteq \mathcal{A}$ s.t. S is a related admissible set and X is a topic of S.

Thus, if an argument does not belong to any admissible set then it does not have an explanation, and an argument has an explanation iff it belongs to an admissible set. As an illustration, all related admissible sets in Example 3 are explanations of A.

To distinguish amongst many different explanations for the same argument, we can classify explanations into two types, as follows:

Definition 4. Given an argument $X \in \mathcal{A}$, let $E_X = \{S | S \text{ is an explanation of } X\}$. Then, for any $S \in E_X$, S is

- a *compact explanation (CE)* for X iff S is smallest, wrt. $\subseteq$, in E_X;
- a *verbose explanation (VE)* for X iff S is largest, wrt. $\subseteq$, in E_X.

Example 4. (Example 3 continued.) $\{A, D, E, F\}$ is a VE. Both $\{A, D, F\}$ and $\{A, E\}$ are CEs. Their natural language reading is:
$\{A, E\}$: choose ic because neither jh nor ritz are cheap;
$\{A, D, F\}$: choose ic as jh is not quiet and ritz is fully booked;
$\{A, D, E, F\}$: choose ic for all reasons above.

3 Computing Explanations

Dispute forests composed of *dispute trees* can be used to compute explanations. A *dispute tree* [2] consists of *proponent* (P) and *opponent* (O) nodes, labelled by arguments (in the given $\langle \mathcal{A}, \mathcal{R} \rangle$); the root is a P node; for every P node, labelled by an argument X, for every Y attacking X, there is a O child of the node, labelled by Y; every O node, labelled by an argument X, has a single P child labelled by an argument attacking X. The set of all P nodes of a dispute tree $\mathcal{T}$ is its *defence set* [2], denoted by $\mathcal{D}(\mathcal{T})$. Dispute trees where no argument labels a P and O node at the same time are called *admissible*: the defence set of an admissible dispute tree is an admissible set [2].

Example 5. (Example 1 continued.) Figure 1 gives four dispute trees for (i.e. with root labelled by) A. All are admissible and there is no other admissible dispute tree for A.

$\mathcal{T}_1$: $[\mathtt{P}:A] \leftarrow [\mathtt{O}:B] \leftarrow [\mathtt{P}:D]$; $[\mathtt{O}:C] \leftarrow [\mathtt{P}:E]$

$\mathcal{T}_2$: $[\mathtt{P}:A] \leftarrow [\mathtt{O}:B] \leftarrow [\mathtt{P}:D]$; $[\mathtt{O}:C] \leftarrow [\mathtt{P}:F]$

$\mathcal{T}_3$: $[\mathtt{P}:A] \leftarrow [\mathtt{O}:B] \leftarrow [\mathtt{P}:E]$; $[\mathtt{O}:C] \leftarrow [\mathtt{P}:E]$

$\mathcal{T}_4$: $[\mathtt{P}:A] \leftarrow [\mathtt{O}:B] \leftarrow [\mathtt{P}:E]$; $[\mathtt{O}:C] \leftarrow [\mathtt{P}:F]$

Figure 1: The four dispute trees for A in Example 1.

Dispute trees are a good match for explanations, as follows:

Theorem 1. Let $X \in \mathcal{A}$:
1. If $\mathcal{T}$ is an admissible dispute tree for X then $S = \mathcal{D}(\mathcal{T})$ is related admissible. Hence S is an explanation for X.
2. If S is an explanation for X, then there is an admissible dispute tree $\mathcal{T}$ s.t. $S' = \mathcal{D}(\mathcal{T})$, $S' \subseteq S$, and S' is admissible.

To support the computation of CEs, we use the following 'more compact than' relation between dispute trees:

Definition 5. For any two dispute trees $\mathcal{T}_i$ and $\mathcal{T}_j$ for the same argument, $\mathcal{T}_i$ is *more compact than* $\mathcal{T}_j$, denoted by $\mathcal{T}_i \prec \mathcal{T}_j$, iff $\mathcal{D}(\mathcal{T}_i) \subset \mathcal{D}(\mathcal{T}_j)$.

In Example 5, $\mathcal{T}_3 \prec \mathcal{T}_1$ and $\mathcal{T}_3 \prec \mathcal{T}_4$. This relation can be applied to trees in a dispute forest, defined as follows:

Definition 6. The *dispute forest for* $X \in \mathcal{A}$ is $\{\mathcal{T} | \mathcal{T} \text{ is an admissible dispute tree for } X\}$.

Thus, a dispute forest is the set of all admissible dispute trees for the same argument. Each tree in a (non-empty) dispute forest for an argument individually justifies it. The dispute trees in Figure 1 form a dispute forest for A.

Theorem 2. Given $X \in \mathcal{A}$, let the dispute forest for X be $\mathcal{F}$ and $\mathcal{T} \in \mathcal{F}$. Furthermore, let $S = \mathcal{D}(\mathcal{T})$. Then S is a CE for X iff $\mathcal{T}$ is smallest, wrt. $\prec$, in $\mathcal{F}$.

In Example 5, $\{A, E\}$ and $\{A, D, F\}$ are (the only) CEs for A, as $\mathcal{T}_2$ and $\mathcal{T}_3$ are (the only) smallest wrt. $\prec$.

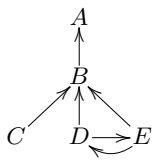
To compute VEs, dispute trees can be grouped into *selected sets*:

Definition 7. Given a dispute forest $\mathcal{F} = \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$, $T \subseteq \mathcal{F}$, $T \neq \{\}$ is a *selected set* (in $\mathcal{F}$) iff for all $\mathcal{T}_i, \mathcal{T}_j \in T$, if $[\mathtt{P}:X]$ is a node in $\mathcal{T}_i$, then $[\mathtt{O}:X]$ is not a node in $\mathcal{T}_j$.

Namely, arguments in defence sets of dispute trees in a selected set do not attack each other, as shown in the following examples.

Example 6. (Example 5 continued.) The selected sets in $\mathcal{F}$ are all non-empty subsets of $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_4\}$.

Example 7. Consider the AA framework in Figure 2 below (top-left) and the dispute forest for A, given by the dispute trees in Figure 2 (bottom). Here, $\mathcal{T}_2$ and $\mathcal{T}_3$ are "incompatible" as D and E label conflicting P/O nodes in these two trees. However, $\mathcal{T}_2$ and $\mathcal{T}_3$ are individually "compatible" with $\mathcal{T}_1$, so the selected sets are: $\{\mathcal{T}_1\}$, $\{\mathcal{T}_2\}$, $\{\mathcal{T}_3\}$, $\{\mathcal{T}_1, \mathcal{T}_2\}$ and $\{\mathcal{T}_1, \mathcal{T}_3\}$.



Admissible sets:
$\{A, C, D\}$, $\{A, C, E\}$, $\{A, C\}$, $\{A, D\}$, $\{A, E\}$, $\{C, D\}$, $\{C, E\}$, $\{C\}$, $\{D\}$, $\{E\}$, $\{\}$
Related admissible sets:
$\{A, C, D\}$, $\{A, C, E\}$, $\{A, C\}$, $\{A, D\}$, $\{A, E\}$, $\{\mathbf{C}\}$, $\{D\}$, $\{E\}$

$\mathcal{T}_1$: $[\mathtt{P}:A] \leftarrow [\mathtt{O}:B] \leftarrow [\mathtt{P}:C]$

$\mathcal{T}_2$: $[\mathtt{P}:A] \leftarrow [\mathtt{O}:B] \leftarrow [\mathtt{P}:D] \leftarrow [\mathtt{O}:E] \leftarrow \ldots$

$\mathcal{T}_3$: $[\mathtt{P}:A] \leftarrow [\mathtt{O}:B] \leftarrow [\mathtt{P}:E] \leftarrow [\mathtt{O}:D] \leftarrow \ldots$

Figure 2

VEs can be computed from selected sets, as follows:

Theorem 3. Given an argument $X \in \mathcal{A}$, let $\mathcal{F}$ be a dispute forest for X, and $T = \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$ be a selected set in $\mathcal{F}$. Let $S_i = \mathcal{D}(\mathcal{T}_i)$ for $i = 1, \ldots, n$ and $S = \bigcup S_i$. Then S is a VE for X iff T is largest, wrt. $\subset$, amongst all selected sets in $\mathcal{F}$.

In Example 6, $\{A, D, E, F\}$ is a VE for A, as $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_4\}$ is the largest selected set. Similarly, in Example 7, $\{A, C, D\}$ and $\{A, C, E\}$ are VEs for A.

4 Conclusion

We have formalised (argumentative) explanations in terms of related admissibility, a restriction over standard admissibility. To help discriminate amongst multiple explanantions for the same argument, we have defined two refined notions of compact and verbose explanations. We have then used dispute trees and forests as a sound and complete computational counterpart for (compact and verbose) explanations. In the future, we plan to study properties of other "related" semantics, e.g., *related grounded-ness*, possibly in other argumentation frameworks, and explore other possible means for explanation comparison.

Acknowledgements

We are grateful to Peter Carruthers for pointing out [4], which has inspired this work. This research was supported by the EPSRC TRaDAr project: EP/J020915/1.

References

[1] P. M. Dung, 'On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games', *AIJ*, **77**(2), 321–357, (1995).
[2] P.M. Dung, P. Mancarella, and F. Toni, 'Computing ideal sceptical argumentation', *AIJ*, **171**(10–15), 642–674, (2007).
[3] X. Fan and F. Toni, 'Decision making with assumption-based argumentation', in *Proc. TAFA*, (2013).
[4] W. H. Newton-Smith, *The Rationality of Science*, Routledge, 1981.