On Computing Explanations in Argumentation

Xiuyi Fan and **Francesca Toni**
{x.fan09,f.toni}@imperial.ac.uk
Department of Computing, Imperial College London, SW7 2AZ, UK

Abstract

Argumentation can be viewed as a process of generating *explanations*. However, existing argumentation semantics are developed for identifying acceptable arguments within a set, rather than giving concrete justifications for them. In this work, we propose a new argumentation semantics, *related admissibility*, designed for giving explanations to arguments in both Abstract Argumentation and Assumption-based Argumentation. We identify different types of explanations defined in terms of the new semantics. We also give a correct computational counterpart for explanations using *dispute forests*.

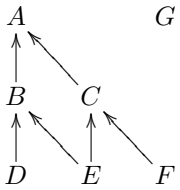
Introduction

Through an arguing process, argumentation has a unique advantage in transparently explaining the procedure and the results of reasoning. Given a topic, the process of arguing can be viewed as identifying *related* information and generating an *explanation* for the topic, usually through some fictitious *proponent* and *opponent* debate game. Hence, arguing for an argument can be deemed to explain it.

Many argumentation semantics have been proposed in the literature. However, existing semantics are designed to answer the question: *Given a set of arguments, which subsets are "good"*? They are less useful in directly answering the question: *Given a set of arguments, why is a particular argument "good"*? Although this question can be answered with "because it belongs to a good set", such answer does not provide a relevant explanation for the argument in question.

We illustrate the motivation of our work with an argumentation-based decision making problem adapted from (Fan and Toni 2013), modelled in Abstract Argumentation (AA) (Dung 1995):

Example 1. An agent needs to decide on accommodation in London, amongst three options: Imperial College Student Accommodation (ic), the John Howard Hotel (jh), and the Ritz Hotel (ritz). The two main criteria for deciding are whether accommodation 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 ritz is fully booked and 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, depicted 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 (neither is jh).
F: Also, ritz is fully booked.
G: London has good public transport.

We can see that D, E and F defend A. Hence, together, they fully justify A. However, E by itself or D and F together also justify A. On the other hand, although G is a piece of valid information, it has nothing to do with A. Hence, if one is interested in explaining A, G should not be included in the explanation.

We propose a new argumentation semantics, *related admissibility*, specifically for generating relevant explanations. We define it in the context of AA as well as Assumption-based Argumentation (ABA) (Dung, Kowalski, and Toni 2009). We choose AA because it is arguably the most widely used argumentation framework, with great simplicity. We choose ABA as a representative of structured argumentation frameworks (Besnard et al. 2014). It is also known that ABA is an instance of AA (Dung, Mancarella, and Toni 2007) and it admits AA as an instance (Toni 2012). Both AA and ABA are well studied with readily usable results.

We identify different types of explanations, all defined in terms of related admissibility. As an illustration, in Example 1, amongst explanations for A, $\{A, E\}$ and $\{A, D, F\}$ are *compact*, as each of the two gives sufficient reasons for A; whereas $\{A, D, F, E\}$ is *verbose*, in that it includes all reasons for A. We use *dispute forests*, composed of dispute trees (Dung, Kowalski, and Toni 2009), as the basis for the computation of (different types of) explanations.

This paper extends (Fan and Toni 2014) in several ways, in particular by considering ABA as well as AA.

Background

Abstract Argumentation (AA) *frameworks* (Dung 1995) are pairs $\langle \mathcal{A}, \mathcal{R} \rangle$, consisting of a set of *abstract arguments*, $\mathcal{A}$, and a binary *attack* relation, $\mathcal{R}$.

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Given an AA framework $AF = \langle \mathcal{A}, \mathcal{R} \rangle$, a set of arguments (or *extension*) $E \subseteq \mathcal{A}$ is *admissible* (in *AF*) iff $\forall A, B \in E$, $(A, B) \notin \mathcal{R}$ (i.e. E is *conflict-free*) and for any $A \in E$, if $(C, A) \in \mathcal{R}$, then there exists some $B \in E$ s.t. $(B, C) \in \mathcal{R}$.

An AA framework can be represented as a directed graph, with arguments being the nodes and attacks being the edges, as we have seen in Example 1.

Assumption-based Argumentation (ABA) *frameworks* (Dung, Kowalski, and Toni 2009; Toni 2014) are tuples $\langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \mathcal{C} \rangle$ [1] where

- $\langle \mathcal{L}, \mathcal{R} \rangle$ is a deductive system, with $\mathcal{L}$ the *language* and $\mathcal{R}$ a set of *rules* of the form $\beta_0 \leftarrow \beta_1, \ldots, \beta_m (m \geq 0, \beta_i \in \mathcal{L})$;
- $\mathcal{A} \subseteq \mathcal{L}$ is a (non-empty) set, referred to as *assumptions*;
- $\mathcal{C}$ is a total mapping from $\mathcal{A}$ into $2^{\mathcal{L}} \setminus \{\{\}\}$,[2] where each $\beta \in \mathcal{C}(\alpha)$ is a *contrary* of α, for $\alpha \in \mathcal{A}$.

Given a rule ρ of the form $\beta_0 \leftarrow \beta_1, \ldots, \beta_m$, β_0 is referred to as the *head* (denoted $Head(\rho) = \beta_0$) and $\beta_1, \ldots, \beta_m$ as the *body* (denoted $Body(\rho) = \{\beta_1, \ldots, \beta_m\}$).

In ABA, *arguments* are deductions of claims using rules and supported by sets of assumptions, and *attacks* are directed at the assumptions in the support of arguments. Informally, following (Dung, Kowalski, and Toni 2009):

- *an argument for (the claim)* $\beta \in \mathcal{L}$ *supported by* $\Delta \subseteq \mathcal{A}$ (denoted $\Delta \vdash \beta$) is a finite tree with nodes labelled by sentences in $\mathcal{L}$ or by τ,[3] the root labelled by β, leaves either τ or assumptions in Δ, and non-leaves β' with, as children, sentences in the body of some rule with head β'.
- An argument $\Delta_1 \vdash \beta_1$ *attacks* an argument $\Delta_2 \vdash \beta_2$ iff β_1 is a contrary of one of the assumptions in Δ_2.

Attacks between (sets of) arguments in ABA correspond to attacks between sets of assumptions, where $\Delta \subseteq \mathcal{A}$ *attacks* $\Delta' \subseteq \mathcal{A}$ iff an argument supported by a subset of Δ attacks an argument supported by a subset of Δ'.

Given $AF = \langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \mathcal{C} \rangle$, *a set of assumptions is admissible* (in AF) iff it does not attack itself and it attacks all $\Delta \subseteq \mathcal{A}$ that attack it; *an argument* $\Delta \vdash \beta$ *is admissible (in* AF*) supported by* $\Delta' \subseteq \mathcal{A}$ iff $\Delta \subseteq \Delta'$ and Δ' is admissible. An argument *is in* AF iff all its rules and assumptions are in AF. $\mathtt{A}^{AF}$ denotes the set of all arguments in AF.

Dispute Trees (Dung, Kowalski, and Toni 2009) are used to prove some of our results. Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$, a *dispute tree for* $A \in \mathcal{A}$ is a (possibly infinite) tree $\mathcal{T}$, s.t.:

1. every node of $\mathcal{T}$ is of the form $[\mathtt{L}\colon X]$, with $L \in \{\mathtt{P}, \mathtt{O}\}$ and $X \in \mathcal{A}$: the node is *labelled* by argument X and assigned the *status* of either *proponent (*$\mathtt{P}$*)* or *opponent (*$\mathtt{O}$*)*;
2. the root of $\mathcal{T}$ is a $\mathtt{P}$ node labelled by A;
3. for every $\mathtt{P}$ node n labelled by an argument B, and for every argument C that attacks B, there exists a child of n, which is an $\mathtt{O}$ node labelled by C;
4. for every $\mathtt{O}$ node n labelled by an argument B, there exists at most one child of n which is a $\mathtt{P}$ node labelled by an argument which attacks B;
5. there are no other nodes in $\mathcal{T}$ except those given by 1-4.

The set of all arguments labelling $\mathtt{P}$ nodes in $\mathcal{T}$ is called the *defence set* of $\mathcal{T}$, denoted by $\mathcal{D}(\mathcal{T})$. A dispute tree $\mathcal{T}$ is an *admissible dispute tree* iff: 1) every $\mathtt{O}$ node in $\mathcal{T}$ has a child, and 2) no argument in $\mathcal{T}$ labels both $\mathtt{P}$ and $\mathtt{O}$ nodes.

Theorem 3.2 in (Dung, Mancarella, and Toni 2007) states: 1) if $\mathcal{T}$ is an admissible dispute tree, then $\mathcal{D}(\mathcal{T})$ is admissible; 2) if $A \in E$ where E is an admissible extension then there exists an admissible dispute tree for A with $\mathcal{D}(\mathcal{T}) = E'$ s.t. $E' \subseteq E$ and E' is admissible.

Explanations in AA

Giving a general theory for explaining human actions/beliefs is a challenging task. It is widely acknowledged that an explanation should be a *justification* (Newton-Smith 1981):

> *...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 AA terms using a 'defends' relation, as follows:

Definition 1. Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$, 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)). Note also that if there is no attack against an argument then its only defence is the argument itself.

Example 2. (Example 1 cntd.) Every argument defends itself. Each of A, D, E and F defends A, and $\{A, D, E, F\}$ and all its non-empty subsets defend A. No argument defends G except G itself.

By combining our 'defends' relation and standard admissibility we obtain our notion of related admissibility:

Definition 2. Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$, $S \subseteq \mathcal{A}$ 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.

Note that, although a self-attacking argument defends itself according to Definition 1, it can never belong to a related admissible set of arguments.

Example 3. (Example 1 cntd.) $\{A, D, E, F\}$, $\{A, D, E\}$, $\{A, D, F\}$, $\{A, E, F\}$, and $\{A, E\}$ are related admissible, with A the topic of all. $\{F, G\}$ is admissible but not related admissible, since F and G do not defend one another.

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

[1] We use $\mathcal{A}$ and $\mathcal{R}$ to represent components of both AA and ABA frameworks, with different meanings. Since we deal with explanations in AA and ABA separately, context will determine which interpretation of these symbols to use.

[2] In some presentations of ABA , contrary maps, equivalently, to single sentences.

[3] $\tau \notin \mathcal{L}$ represents "true" and stands for the empty body of a rule.

Proposition 1. Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$ and 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 thereof whose topic is D is $\{D\}$.

We use related admissible sets to define explanations:

Definition 3. Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$, 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.

Since we define explanation as a set of arguments, we can characterise explanations in terms of relations between sets. We will use the following relation:

Definition 4. Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$ and an argument $A \in \mathcal{A}$, let S_i, S_j be explanations of A. Then S_i is *smaller than* S_j, denoted by $S_i < S_j$, iff $|S_i| < |S_j|$.[4]

We can classify explanations into different types:

Definition 5. Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$, let $A \in \mathcal{A}$ and $E_A = \{S | S$ is an explanation of $A\}$. Then, for any $S \in E_A$, S is

- a *Minimal Explanation (MiE)* iff S is smallest wrt $<$;
- a *Compact Explanation (CE)* iff S is smallest wrt $\subset$;
- a *Maximal Explanation (MaE)* iff S is largest wrt $<$;
- a *Verbose Explanation (VE)* iff S is largest wrt $\subset$.

Intuitively, MiEs and CEs are succinct whereas MaEs and VEs are comprehensive. The following result trivially holds.

Proposition 2. Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$, for any argument $A \in \mathcal{A}$ and explanation S of A, if S is a MaE, then S is also a VE; if S is a MiE, then S is also a CE.

Example 4. (Example 3 cntd.) $\{A, D, E, F\}$ is both a MaE and a VE. Both $\{A, D, F\}$ and $\{A, E\}$ are CEs. $\{A, E\}$ is a MiE. Their natural language reading is:

$\{A, E\}$: choose ic as both jh and ritz are not cheap.
$\{A, D, F\}$: choose ic as jh is not quiet and ritz is booked.
$\{A, D, E, F\}$: choose ic for all reasons above.

Computing Explanations in AA

Here, we show the computation of explanations with *dispute forests* composed of dispute trees. Admissible dispute trees correspond to admissible arguments (Dung, Mancarella, and Toni 2007). Hence, given an admissible dispute tree $\mathcal{T}$, related admissible sets of arguments can be extracted from $\mathcal{T}$.

Theorem 1. Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$ and an argument $X \in \mathcal{A}$, let $\mathcal{T}$ be a dispute tree for X.

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

[4]For a set S, $|S|$ denotes the cardinality of S.

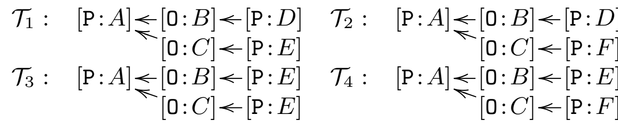


Figure 1: Dispute trees for A in Example 1 (see Example 5).

Proof. (Sketch.) Since all P nodes in a dispute tree defend the root, both directions of this theorem hold by Theorem 3.2 in (Dung, Mancarella, and Toni 2007). □

Example 5. (Example 3 cntd.) The four dispute trees for A, shown in Figure 1, are all admissible dispute trees. There is no other admissible dispute tree for A. The defence sets of these trees are all explanations, and each explanation has one of the defence sets as a subset.

Admissible dispute trees form dispute forest, as follows.

Definition 6. Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$ and an argument $A \in \mathcal{A}$, the *dispute forest for* A is $\{\mathcal{T} | \mathcal{T}$ is an admissible dispute tree for $A\}$.

Thus, a dispute forest is the set of all admissible dispute trees for the same argument. Each tree individually gives the justification for its root, as seen in Figure 1 for Example 1.

To compute explanations of different types, we define *smaller* and *more compact* relations between dispute trees.

Definition 7. For any two dispute trees $\mathcal{T}_i$ and $\mathcal{T}_j$ for the same argument, $\mathcal{T}_i$ is *smaller than* $\mathcal{T}_j$, denoted by $\mathcal{T}_i < \mathcal{T}_j$, iff $|\mathcal{D}(\mathcal{T}_i)| < |\mathcal{D}(\mathcal{T}_j)|$. $\mathcal{T}_i$ is *more compact than* $\mathcal{T}_j$, denoted by $\mathcal{T}_i \prec \mathcal{T}_j$, iff $\mathcal{D}(\mathcal{T}_i) \subset \mathcal{D}(\mathcal{T}_j)$.

Example 6. (Example 5 cntd.) For the trees in Figure 1, we have $\mathcal{T}_3 < \mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_4$ and $\mathcal{T}_3 \prec \mathcal{T}_1, \mathcal{T}_4$.

With the above relations defined over dispute trees, MiEs and CEs can be obtained from dispute forests:

Theorem 2. Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$ and $A \in \mathcal{A}$, let the dispute forest for A be $\mathcal{F}$ and $\mathcal{T} \in \mathcal{F}$. Furthermore, let $S = \mathcal{D}(\mathcal{T})$. Then, S is a MiE for A iff $\mathcal{T}$ is smallest wrt $<$ in $\mathcal{F}$; S is a CE for A iff $\mathcal{T}$ is smallest wrt $\prec$ in $\mathcal{F}$.

Proof. (Sketch.) We prove that S is a MiE iff $\mathcal{T}$ is smallest wrt $<$. The proof for CE is similar. By Theorem 1, we know that S is an explanation as $\mathcal{T}$ is admissible. Moreover, there is no other explanation not captured by trees in $\mathcal{F}$. Since $\mathcal{T}$ is smallest, by Definition 7, S is smallest. Conversely, if S is a MiE, then by Theorem 1, there is a $\mathcal{T}$ in $\mathcal{F}$ s.t. $S' = \mathcal{D}(\mathcal{T})$. Since S is smallest, $S = S'$ and $\mathcal{T}$ is smallest in $\mathcal{F}$. □

Example 7. (Example 6 cntd.) For A, $\{A, E\}$ is a MiE, as $\mathcal{T}_3$ is smallest wrt $<$; both $\{A, E\}$ and $\{A, D, F\}$ are CEs, as both $\mathcal{T}_3$ and $\mathcal{T}_2$ are smallest wrt $\prec$.

To compute MaEs and VEs, dispute trees are grouped into *selected sets*, as follows.

Definition 8. 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 $[\mathtt{P}: B]$ is a node in $\mathcal{T}_i$, then $[\mathtt{O}: B]$ is not a node in $\mathcal{T}_j$.

For any selected set S, arguments in defence sets of debate trees from S do not attack each other, as shown in the following two examples.

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

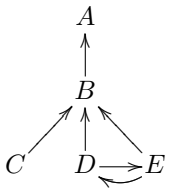
$\mathcal{T}_2$: $[\mathtt{P}:A] \longleftarrow [\mathtt{O}:B] \longleftarrow [\mathtt{P}:D] \longleftarrow [\mathtt{O}:E] \longleftarrow \ldots$

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

Figure 2: The dispute forest for A with dispute trees: $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$ (see Example 9).

Example 8. (Example 5 cntd.) 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 9. An AA framework is given below. The dispute forest for A is in Figure 2. Here, $\mathcal{T}_2$ and $\mathcal{T}_3$ are "incompatible" as D and E have conflicting P/O status in these trees, but $\mathcal{T}_2$ 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:
$\{A, C, D\}$, $\{A, C, E\}$, $\{A, C\}$, $\{A, D\}$, $\{A, E\}$, $\{C\}$, $\{D\}$, $\{E\}$

We can also compare selected sets as follows.

Definition 9. Given two selected sets $T_i = \{\mathcal{T}_1^i, \ldots, \mathcal{T}_n^i\}, T_j = \{\mathcal{T}_1^j, \ldots, \mathcal{T}_m^j\}$, let $S_k^i = \mathcal{D}(\mathcal{T}_k^i)$ and $S_l^j = \mathcal{D}(\mathcal{T}_l^j)$, for $k = 1, \ldots, n$ and $l = 1, \ldots, m$. We say that T_i is *smaller than* T_j, denoted by $T_i < T_j$ iff $|S_i| < |S_j|$ where $S_i = \bigcup S_k^i$ and $S_j = \bigcup S_l^j$. T_i is *more compact than* T_j, denoted by $T_i \prec T_j$, iff $S_i \subset S_j$.

MaEs and VEs can be computed from selected sets.

Theorem 3. Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$ and an argument $A \in \mathcal{A}$, let $\mathcal{F}$ be a debate forest for A, 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 MaE for A iff T is a largest selected set in $\mathcal{F}$ wrt $<$; and S is a VE for A iff T is a largest selected set in $\mathcal{F}$ wrt $\prec$.

Proof. (Sketch.) We prove that S is a MaE iff T is largest wrt $<$ The proof for VEs is similar. We first prove that if T is largest then S is a MaE. By Theorem 1, we know that S_i is admissible. By Definition 8, S is conflict-free. Hence, we can see that S is admissible. Since all arguments in S defend A, S is an explanation of A. By Definition 9, we can see that S is a MaE. If S is a MaE, since S is an explanation, then S is conflict-free and all arguments in S defend A. Hence, we can see that S is formed by extracting all proponent nodes from a selected set T. By Definition 9, if S is largest, T is largest as well. □

We illustrate Theorem 3 with the following two examples.

Example 10. (Example 7 cntd.) For topic argument A, $\{A, D, E, F\}$ is both a MaE and VE as $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_4\}$ is largest wrt both $<$ and $\subset$.

Example 11. (Example 9 cntd.) For topic argument A, $\{A, C, D\}$ and $\{A, C, E\}$ are both MaEs and VEs as $\{\mathcal{T}_1, \mathcal{T}_2\}$ and $\{\mathcal{T}_1, \mathcal{T}_3\}$ are both largest wrt $<$ and $\prec$. Also $\{A, C\}$, $\{A, D\}$, and $\{A, E\}$ are MiEs and CEs as $\{\mathcal{T}_1\}$, $\{\mathcal{T}_2\}$ and $\{\mathcal{T}_3\}$ are both smallest wrt $<$ and $\prec$.

Explanations in ABA and their Computation

We have introduced explanations in AA and discussed their computation via dispute trees and forests. In this section, we extend their use in ABA. The following example provides motivation and illustration grounds.

Example 12. We revise Example 1 such that a third decision criterion *near* is introduced and we consider only two choices, John Howard Hotel (jh) and Imperial College Student Accommodation (ic). We let both jh and ic be near (ic remains cheap and quiet whereas jh is neither cheap nor quiet). To model this decision problem we use an ABA framework adapted from the *Weakly Dominant Decision Framework* in (Fan and Toni 2013), such that a selected decision meets at least one goal not met by others. The ABA framework $AF = \langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \mathcal{C} \rangle$ consists of:[5]

$\mathcal{R}$: $\{$ $more(X, Z) \leftarrow met(X, Y), nMet(Z, Y)$;
$nSel(X) \leftarrow met(Z, Y), nMet(X, Y), notMore(X, Z)$;
$met(ic, cheap) \leftarrow$; $met(ic, quiet) \leftarrow$; $met(ic, near) \leftarrow$;
$met(jh, near) \leftarrow\}$;
$\mathcal{A}$: $\{$ $sel(ic)$; $sel(jh)$; $notMore(ic, jh)$; $notMore(jh, ic)$;
$nMet(ic, cheap)$; $nMet(ic, quiet)$; $nMet(ic, near)$;
$nMet(jh, cheap)$; $nMet(jh, quiet)$; $nMet(jh, near)\}$;
$\mathcal{C}$: $\mathcal{C}(notMore(X, Z)) = \{more(X, Z)\}$;
$\mathcal{C}(sel(X)) = \{nSel(X)\}$; $\mathcal{C}(nMet(X, Y)) = \{met(X, Y)\}$.

This ABA framework can be interpreted as follows. By default, each choice is assumed to be a good decision, so both $sel(ic)$ and $sel(jh)$ are assumptions (standing for "select ic" and "select jh", resp). The contraries are $nSel(ic)$ and $nSel(jh)$ ("not to select ic" and "not to select jh"), resp. Criteria met by each choice are specified with rules with empty body, e.g., ic meets the criterion cheap, hence having the rule $met(ic, cheap) \leftarrow$. A choice X is not a good decision if it does not meet a criterion Y yet some other decision Z meets Y; and it is not the case that X meets some criterion not met by Z. This condition is expressed with:

$$nSel(X) \leftarrow met(Z, Y), nMet(X, Y), notMore(X, Z).$$

The condition that X meets some criterion Y not met by Z is specified by the rule:

$$more(X, Z) \leftarrow met(X, Y), nMet(Z, Y).$$

Arguments in *AF* include:
$A = \{sel(ic)\} \vdash sel(ic)$, $B = \{\} \vdash met(ic, near)$,
$C = \{nMet(jh, cheap)\} \vdash more(ic, jh)$,
$D = \{nMet(jh, quiet)\} \vdash more(ic, jh)$, and
$E = \{nMet(ic, near), notMore(ic, jh)\} \vdash nSel(ic)$.
We can see that E attacks A. Also, B, C, and D attack E. Nothing attacks B, C or D hence $\{A, B\}$, $\{A, C\}$, $\{A, D\}$, $\{A, B, C\}$, $\{A, B, D\}$, $\{A, C, D\}$ and $\{A, B, C, D\}$ are admissible.

The *defends* relation defined for AA (Definition 1) can also be used for ABA. In ABA though we can also define a *defends* relation between sentences and arguments, as follows:

[5]Rule/assumptions/contrary schemata (with variables X, Y, Z) are used to stand for the set of all their instances wrt constants (ic and jh for X, Z and $cheap$, $quiet$, $near$ for Y).

Definition 10. Given an ABA framework $\langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \mathcal{C} \rangle$, let $\chi \in \mathcal{L}$ and $A, B \in \mathtt{A}^{AF}$. Then A *defends* χ iff χ is the claim of B and A defends B.

Note that, for any argument $A = \_ \vdash \chi$, A defends χ.[6]

Example 13. (Example 12 cntd.) Arguments A, B, C and D defend $sel(ic)$.

We define related admissibility in ABA, as follows.

Definition 11. Given $AF = \langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \mathcal{C} \rangle$, a set of arguments $S \subseteq \mathtt{A}^{AF}$ is *related admissible* iff:

1. S is admissible,
2. there exists a *topic* sentence χ (of S) s.t. χ is the claim of some argument in S and for all $B \in S$, B defends χ.

Definition 11 is as Definition 2, but instead of letting all arguments in a related admissible set defend a topic argument, they defend the claim of a topic argument.

Example 14. (Example 13 cntd.) For the topic sentence $sel(ic)$, $\{A, B\}, \{A, C\}$, $\{A, D\}$, $\{A, B, C\}$, $\{A, B, D\}$, $\{A, C, D\}$ and $\{A, B, C, D\}$ are related admissible.

As a structured argumentation formalism, ABA allows analysing arguments at a fine-grained level, in terms of rules, assumptions and contraries. We show that related admissibility can also be defined with assumptions. We start with defining the *defends relation* between the topic of an argument and assumptions as follows:

Definition 12. Given $AF = \langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \mathcal{C} \rangle$, let $\chi \in \mathcal{L}$ and $\alpha \in \mathcal{A}$. Then α *defends* χ iff there exists $B = \Delta \vdash \_, \alpha \in \Delta$ and $A = \_ \vdash \chi$ in $\mathtt{A}^{AF}$ s.t. B defends A.

From this definition, we can see that given an argument $A = \Delta \vdash \chi$, all assumptions in Δ defend χ, and all assumptions in arguments defending A defend χ.

Example 15. With AF shown in Example 12, assumptions $sel(ic)$, $nMet(jh, quiet)$ and $nMet(jh, cheap)$ defend $sel(ic)$.

Definition 13. Given $\langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \mathcal{C} \rangle$, a set of assumptions $\Delta \subseteq \mathcal{A}$ is *related admissible* iff:

1. Δ is admissible,
2. there exists a *topic* sentence χ (of Δ), $\chi \in \mathcal{L}$, s.t. for all assumptions $\alpha \in \Delta$, α defends χ.

Example 16. (Example 15 cntd.) $\{sel(ic)\}$, $\{sel(ic), nMet(jh, cheap)\}$, $\{sel(ic), nMet(jh, quiet)\}$ and $\{sel(ic), nMet(jh, cheap), nMet(jh, quiet)\}$ are related admissible as they are all admissible sets of assumptions including the topic sentence $sel(ic)$.

As shown by Examples 12 and 16, the two views of related admissibility wrt arguments and assumptions given by Definition 11 and 13, resp, coincide, as follows:

Theorem 4. Given an ABA framework AF, if a set of arguments $S \subseteq \mathtt{A}^{AF}$ is related admissible then the set of assumption $\Delta = \cup_{\Delta' \vdash \_ \in S} \Delta'$ is related admissible.

[6]Throughout, _ stands for an anonymous variable.

$$\mathcal{T}_1: \quad \begin{array}{c} [\mathtt{P}: \{sel(ic)\} \vdash sel(ic)] \\ \uparrow \\ [\mathtt{O}: \{nMet(ic, near), notMore(ic, jh)\} \vdash nSel(ic)] \\ \uparrow \\ [\mathtt{P}: \{\} \vdash met(ic, near)] \end{array}$$

$$\mathcal{T}_2: \quad \begin{array}{c} [\mathtt{P}: \{sel(ic)\} \vdash sel(ic)] \\ \uparrow \\ [\mathtt{O}: \{nMet(ic, near), notMore(ic, jh)\} \vdash nSel(ic)] \\ \uparrow \\ [\mathtt{P}: \{nMet(jh, cheap)\} \vdash more(ic, jh)] \end{array}$$

$$\mathcal{T}_3: \quad \begin{array}{c} [\mathtt{P}: \{sel(ic)\} \vdash sel(ic)] \\ \uparrow \\ [\mathtt{O}: \{nMet(ic, near), notMore(ic, jh)\} \vdash nSel(ic)] \\ \uparrow \\ [\mathtt{P}: \{nMet(jh, quiet)\} \vdash more(ic, jh)] \end{array}$$

Figure 3: The dispute trees for $sel(ic)$ in Example 12.

Proof. (Sketch.) Δ is admissible as S is admissible. The topic χ of S is the topic of Δ and all $\alpha \in \Delta$ defend χ as there is some $A \in S$ s.t. α is an assumption in A, since A defends χ, so α does. Thus Δ is related admissible. □

Definition 5 has introduced MiE, CE, MaE and VE for AA. These definitions apply in ABA as well as it is an instance of AA. We do not repeat them. Since, in general, there is no correspondence between the number of arguments in an explanation and the number of assumptions contained in these arguments, we cannot conclude that a MiE has minimal number of assumptions or a MaE has maximal number of assumptions, as illustrated in the following example.

Example 17. Given an ABA framework $AF = \langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \mathcal{C} \rangle$ with $\mathcal{R} = \{d \leftarrow; f \leftarrow\}$, $\mathcal{A} = \{a, b, c, e\}$, $\mathcal{C}(a) = \{b, c\}, \mathcal{C}(b) = \{d, e\}, \mathcal{C}(c) = \{e, f\}, \mathcal{C}(e) = \{z\}$, the following arguments are in $\mathtt{A}^{AF}$:

$A = \{a\} \vdash a$, $B = \{b\} \vdash b$, $C = \{c\} \vdash c$, $E = \{e\} \vdash e$, $D = \{\} \vdash d$, $F = \{\} \vdash f$.

Attacks are as shown in Example 1 (except that there is no argument G here). Explanations for A are $\{A, E\}$, $\{A, D, F\}$ and $\{A, E, D, F\}$. Amongst them, $\{A, E\}$ is a MiE and $\{A, D, F\}$ is not a MiE. However, there is only one assumption in $\{A, D, F\}$: a; yet there are two assumptions in $\{A, E\}$: a, e. Hence, in this example, a MiE does not have the minimum number of assumptions. Similar examples can be given for MaE as well.

To find related admissible sets of arguments and assumptions, we again use dispute trees and forests. We do not repeat the two definitions, but let a dispute tree for ABA be defined for a topic sentences χ rather than an argument. Also, in ABA, we let the root of a debate tree for χ be an argument with claim χ, illustrated below.

Example 18. (Example 16 cntd.) Given the AF shown in Example 12, three dispute trees, $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$ for $sel(ic)$ are shown in Figure 3. They are all admissible. Hence, ic should be selected because: ($\mathcal{T}_1$) ic is near; ($\mathcal{T}_2$) ic is cheap whereas jh is not; and ($\mathcal{T}_3$) ic is quiet whereas jh is not. Each tree gives a different reason for selecting ic over jh.

Dispute trees can be used to compute related admissible assumptions.

Theorem 5. Given an ABA framework $\langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \mathcal{C} \rangle$ and a sentence $\chi \in \mathcal{L}$, let $\mathcal{T}$ be an admissible dispute tree for χ. Then $\{\alpha | [\mathtt{P} : \Delta \vdash \_] \text{ is in } \mathcal{T} \text{ and } \alpha \in \Delta\}$ is related admissible.

The proof of this theorem is straightforward from Definition 11 and Theorem 4.

The concept of dispute forest as given in Definition 6 also applies in ABA, with the modification that a dispute forest in ABA is wrt a sentence, as follows.

Definition 14. Given an ABA framework $\langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \mathcal{C} \rangle$ and a sentence $\chi \in \mathcal{L}$, the dispute forest for χ is $\{\mathcal{T} | \mathcal{T}$ is an admissible dispute tree for $\chi\}$.

The three dispute trees in Figure 3 for $sel(ic)$ form the dispute forest for $sel(ic)$ in *AF* shown in Example 12. We do not need to redefine selected set (Definition 8) for ABA. It is easy to see that the equivalent of Theorems 2 and 3 hold for ABA, illustrated with the following example.

Example 19. (Example 18 cntd.) Let A, B, C, D be as defined in Example 12, then $\{A, B\}$, $\{A, C\}$, $\{A, D\}$ are MiEs and CEs. $\{A, B, C, D\}$ is a MaE and a VE.

Related Work

(Moulin et al. 2002) survey explanation capabilities of knowledge-based systems and decision support systems extensively, give a philosophical account of many developed formalisms and applications. Though argumentation is considered as a means for explanation, no work surveyed there was dedicated to understanding explanation as a formal computational argumentation semantics.

(García et al. 2013) study dialectical explanation for argument-based reasoning in knowledge-based systems. Differences between that work and ours include that i) they view explanations as sets of trees whereas we define explanations as semantics; ii) they rely on labelling as in (García and Simari 2004) for computing explanations whereas we use dispute trees; iii) for structured argumentation, they exemplify their notions in DELP (García and Simari 2004) whereas we use ABA. It can be argued that trees are a more comprehensive representation than sets of arguments. Yet, we believe that our treatment of explanation as semantics gives a more direct answer to the question we set to answer: "why an argument is accepted."

(Lacave and Diez 2004) review explanation methods for expert systems. As noted in (Moulin et al. 2002), argumentation is used as a means for explanation in some expert systems. These are not concerned about explaining the acceptability of arguments in argumentation as we do.

(Dung, Kowalski, and Toni 2006) present a dialectic proof procedure for admissibility semantics in ABA using dispute trees. Our work presents a new semantics based on admissibility that is suitable for both AA and ABA and can be computed with dispute forests. The emphasis on generating explanations for arguments has not been previously studied.

(Dung, Toni, and Mancarella 2010) propose three principles as a guideline for designing argumentation systems: (1) Arguments must be simple (Transparency); (2) Arguments must be given in full (Relevance); (3) All counter-arguments must be considered (No dismissal). We can draw analogy from this list to argumentative explanations. It would be interesting to see whether our explanation types are linked to these principles, since each explanation contains no unrelated argument and all defences for counter-arguments are included in an explanation.

(Baroni and Giacomin 2007) give semantics evaluation criteria. Considering explanations as semantics, MaE and VE meet their I-maximality, admissibility and directionality, whereas MiE and CE meet admissibility and directionality.

(Baroni and Giacomin 2007) also introduce the notion of strong admissibility. Comparing with their notion, the differences are: 1) our work is motivated by giving explanations to arguments, not purely semantics modifications; 2) we only need the notion of defense, rather than strong defense as needed for strong admissibility. Thus a relevant admissible argument is allowed to defend itself; and (3) strong admissibility does not give relevance, i.e., two disjoint groups of arguments can be in a single strong admissible set, e.g., given an abstract argumentation framework with two arguments A, B and no attack; then the set $\{A, B\}$ is strongly admissible. Yet, $\{A, B\}$ is not relevant admissible.

(Thang, Dung, and Hung 2009) give a framework for computing grounded, ideal and preferred semantics using debate trees. They are concerned about defining a generic proof procedure for several semantics whereas we focus on explanation; they use base derivation to track multiple ways of defending the topic whereas we use debate forests.

(Craven, Toni, and Williams 2013) give a graph based dispute derivation method to compute admissible, complete and grounded semantics of ABA. They introduce rule-minimal argument to improve the efficiency of the semantics computation. Their process does not necessarily find MiEs or CEs for a topic argument but may find smaller explanations in general. That work is not motivated by computing explanations, but rather computational efficiency.

(Schulz and Toni 2013) presents a work on using ABA to explain why a literal is (or not) contained in an answer set of a logic program. Their work uses ABA as the tool for explanation with the stable semantics (Dung 1995; Bondarenko et al. 1997) whereas we focus on how to explain arguments in AA and ABA. It would be interesting to see whether our method can be applied in their work and help them give better explanations.

(Zhong et al. 2014) present a work on applying argumentation-based decision making in a legal application. They present an algorithm to generate natural language explanations from debate trees. Their algorithm is domain specific and solely concerns admissibility. It would be interesting to see that if related admissibility is a more suitable semantics for their applications and our results allow for better natural language explanations to be constructed.

Conclusion

In this work, we formalise the concept of argumentative explanation as a novel argumentation semantics, related admissibility. We aim at directly answering the question: *Why*

is an argument A accepted in an argumentation framework? We let an explanation of A be a set of arguments S justifying A. Given by the related admissibility semantics, S only contains arguments defending A. Since multiple explanations can be given to an argument, we define different types of explanations and discuss their computation.

We present the related admissibility semantics in both AA and ABA. The computation is discussed in the form of dispute trees and forests. We have shown that our approach is sound and complete. The contribution includes: (1) formalisation of several notions of argumentative explanations and (2) introduction of a new argumentation semantics with its computation for both AA and ABA.

In the future, we would like to study properties of other "related" semantics, e.g., *related grounded* or *related ideal*. This work uses debate trees and forests for semantics computation. Labelling (Modgil and Caminada 2009; Caminada and Gabbay 2009) is another approach for semantics computation in AA. It would be interesting to see if any of their methods apply in our work. Also, this work has studied explanation at the argument level. As we illustrated in Example 17, there is a disconnection between the number of arguments in an explanation and the number of assumptions therein. It would be interesting to further explore the concept of explanation at the level of assumptions or rules. Moreover, we would like to see if our method applies in other structured argumentation frameworks (Besnard et al. 2014), such as ASPIC+ (Prakken 2010) and Logical Argumentation (Besnard and Hunter 2001). Lastly, this work is solely about explaining why arguments hold but not why arguments do not hold. It would be interesting to see if our method can be adapted to answer the negative side of the question as well.

Acknowledgements

We are grateful to P. Carruthers for pointing out (Newton-Smith 1981), which has inspired this work. This research was supported by the EPSRC TRaDAr project *Transparent Rational Decisions by Argumentation*: EP/J020915/1.

References

Baroni, P., and Giacomin, M. 2007. On principle-based evaluation of extension-based argumentation semantics. *AIJ* 171(10-15):675–700.

Besnard, P., and Hunter, A. 2001. A logic-based theory of deductive arguments. *AIJ* 128(1-2):203–235.

Besnard, P.; Garcia, A.; Hunter, A.; Modgil, S.; Prakken, H.; Simari, G.; and Toni, F. 2014. *Argument & Computation, Special Issue: Tutorials on Structured Argumentation* 5(1).

Bondarenko, A.; Dung, P.; Kowalski, R.; and Toni, F. 1997. An abstract, argumentation-theoretic approach to default reasoning. *AIJ* 93(1-2):63–101.

Caminada, M., and Gabbay, D. 2009. A Logical Account of Formal Argumentation. *Studia Logica* 93(2):109–145.

Craven, R.; Toni, F.; and Williams, M. 2013. Graph-based dispute derivations in assumption-based argumentation. In *Proc. TAFA*, 46–62. Springer.

Dung, P.; Kowalski, R.; and Toni, F. 2006. Dialectic proof procedures for assumption-based, admissible argumentation. *AIJ* 170:114–159.

Dung, P.; Kowalski, R.; and Toni, F. 2009. Assumption-based argumentation. In *Arg. in AI*. Springer. 199–218.

Dung, P.; Mancarella, P.; and Toni, F. 2007. Computing ideal sceptical argumentation. *AIJ* 171(10–15):642–674.

Dung, P.; Toni, F.; and Mancarella, P. 2010. Some design guidelines for practical argumentation systems. In *Proc. COMMA*, 171–182. IOS Press.

Dung, P. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *AIJ* 77(2):321–357.

Fan, X., and Toni, F. 2013. Decision making with ABA. In *Proc. TAFA*, 127–142. Springer.

Fan, X., and Toni, F. 2014. On computing explanation in abstract argumentation. In *Proc. ECAI*.

García, A. J., and Simari, G. R. 2004. Defeasible logic programming an argumentative approach. In *TPLP*, 95–138. Cambridge University Press.

García, A. J.; Chesñevar, C. I.; Rotstein, N. D.; and Simari, G. R. 2013. Formalizing dialectical explanation support for argument-based reasoning in knowledge-based systems. *Expert Systems with Applications* 40(8):3233 – 3247.

Lacave, C., and Diez, F. J. 2004. A review of explanation methods for heuristic expert systems. *The Knowledge Engineering Review* 19:133–146.

Modgil, S., and Caminada, M. 2009. Proof theories and algorithms for abstract argumentation frameworks. In Rahwan, I., and Simari, G., eds., *Argumentation in AI*. Springer.

Moulin, B.; Irandoust, H.; Blanger, M.; and Desbordes, G. 2002. Explanation and argumentation capabilities:towards the creation of more persuasive agents. *AI Review* 17(3):169–222.

Newton-Smith, W. H. 1981. *The Rationality of Science*. Routledge.

Prakken, H. 2010. An abstract framework for argumentation with structured arguments. *Arg. & Comp.* 1(2):93–124.

Schulz, C., and Toni, F. 2013. ABA-based answer set justification. *TPLP* 13(4-5-Online-Supplement).

Thang, P.; Dung, P.; and Hung, N. 2009. Towards a common framework for dialectical proof procedures in abstract argumentation. *JLC* 19(6):1071–1109.

Toni, F. 2012. Reasoning on the web with assumption-based argumentation. In *Reasoning Web. Semantic Technologies for Advanced Query Answering*, volume 7487. Springer. 370–386.

Toni, F. 2014. A tutorial on assumption-based argumentation. *Argument & Computation, Special Issue: Tutorials on Structured Argumentation* 5(1):89–117.

Zhong, Q.; Fan, X.; Toni, F.; and Luo, X. 2014. Explaining best decisions via argumentation. In *Proc. ECSI*, 224–237. CEUR.