Computing Argumentative Explanations in Bipolar Argumentation Frameworks

Zhiwei Zeng,[1,2,3] **Chunyan Miao,**[1,3,4] **Cyril Leung,**[1,6] **Zhiqi Shen,**[1,4] **Jing Jih Chin**[5]
[1]LILY Research Center, [2]IGS, [3]Alibaba-NTU Singapore JRI, [4]SCSE, [5]LKC School of Medicine,
Nanyang Technological University, Singapore
[6]ECE, University of British Columbia, Vancouver, Canada
i160001@e.ntu.edu.sg

Introduction

The process of arguing is also the process of justifying and explaining. Transparent reasoning process endows argumentation good explainability. Recently, more research efforts have been devoted to realizing the explanatory power of argumentation (Fan and Toni 2015; Zeng et al. 2018) in unipolar argumentation frameworks. In addition to the attack relation, bipolar frameworks consider the support relation, which brings greater expressibility but also complexity. It is worth exploring how the interactions encompassed in the support relation contribute to the arguing process and how to capture them in explanations. In this paper, we propose a "stronger" notion of defence and a new *bipolar admissibility* semantics, which are defined based on both the attack and the support relations, and use them to formalize two types of explanations, namely *concise* and *strong* explanations. We then present complete and sound processes for computing explanations by constructing *bipolar dispute trees*.

An **Abstract Bipolar Argumentation Framework (BAF)** (Cayrol and Lagasquie-Schiex 2005) is a tuple $BAF = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+ \rangle$, consisting of a set of arguments $\mathcal{A}$, a binary *attack* relation $\mathcal{R}^-$ on $\mathcal{A}$, and another binary *support* relation $\mathcal{R}^+$ on $\mathcal{A}$.

Let $a, b \in \mathcal{A}$, a *supports* b iff there exists a sequence $(a_1, \cdots, a_n)$ of elements of $\mathcal{A}$, such that $n \geq 2, a = a_1, b = a_n$, $a_1\mathcal{R}^+a_2, \cdots, a_{n-1}\mathcal{R}^+a_n$. Argument a forms a *supported attack* on b iff there exists a sequence (a, x, b) of arguments such that a supports x and $x\mathcal{R}^-b$.

Let $S \subseteq \mathcal{A}$, $a \in \mathcal{A}$, S *set-attacks* a iff $\exists b \in S$ such that b forms a direct or supported attack on a. S *set-supports* a iff $\exists b \in S$ such that b supports a.

Let $S \subseteq \mathcal{A}$, S is *conflict-free+* iff $\nexists a, b \in S$ such that $\{a\}$ set-attacks b. S is *safe* iff S is conflict-free+ and $\nexists b \in \mathcal{A}$ such that both S set-attacks b and S set-supports b. If S is safe, then S is conflict-free+. If S is conflict-free+ and closed for $\mathcal{R}^+$, then S is safe (Cayrol and Lagasquie-Schiex 2005).

Explanations in BAF

A set of arguments that contributes to the justification of an argument a by defending a and its defenders can be used for explaining a. Guided by this idea, we first define two types of defence in BAF, with and without reference to the support relation respectively.

Definition 1. Given a BAF framework $\langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+ \rangle$, let $S \subseteq \mathcal{A}, a \in \mathcal{A}$. S *defends* a iff $\forall b \in \mathcal{A}$, if $b\mathcal{R}^-a$, then $\exists c \in S$ such that $c\mathcal{R}^-b$.

Definition 2. A set $S \subseteq \mathcal{A}$ *strongly defends* argument $a \in \mathcal{A}$ iff $\forall b \in \mathcal{A}$, if $\{b\}$ set-attacks a, then $\exists c \in S$ such that $c\mathcal{R}^-b$.

Example 1. As shown in Fig. 1(1), argument b directly attacks argument a. The sequence (c, b, a), (d, b, a) of arguments are two supported attacks on argument a by c and d respectively. $\{e\}$ defends a as $e\mathcal{R}^-b$. Set $\{e, f, g\}$ and $\{e, f, g, h, i, j\}$ strongly defend a. However, set $\{e\}$ and $\{f, g, h\}$ do not strongly defend a, as they cannot directly counter-attack all direct and supported attacks on a.

Using the notion of set-attack, Definition 2 gives a new form of defence which is more stringent than the traditional Dung's style defence re-contexted in Definition 1. For a set S to strongly defend an argument a, S needs to counter not only all direct attacks, but also all supported attacks on a.

With the two notions of defence, we then propose two types of admissibility. The *u-admissibility* (unipolar admissibility) inherits Dung's definition of admissibility in Abstract Argumentation (AA). The *b-admissibility* (bipolar admissibility) is constructed based on the definition of "strongly defend" and enforces external coherence.

Definition 3. Given a set of arguments $S \subseteq \mathcal{A}$:
- S is *u-admissible* iff S is conflict-free+ and defends all its elements.
- S is *b-admissible* iff S is safe and strongly defends all its elements.

Proposition 1. *Given a BAF framework $\langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+ \rangle$, let $S \subseteq \mathcal{A}$. If S is b-admissible, then S is also u-admissible, but not vice versa.*

Example 2. (cont'd) $\{a, e, f, g, l\}$ and $\{i, k, l, h\}$ are b-admissible and u-admissible. $\{a, e\}$, $\{a, e, f, h\}$ (a not strongly defended) are u-admissible, but not b-admissible.

Definition 4. Let $a, b \in \mathcal{A}, S \subseteq \mathcal{A}$. a is *relevant* to b iff there exists a sequence $(a_1, ..., a_n)$ of elements of $\mathcal{A}$ such that $n \geq 2$, $a = a_1$, $b = a_n$, $a_1\mathcal{R}a_2, \cdots, a_{n-1}\mathcal{R}a_n$, with $\mathcal{R} \in \{\mathcal{R}^-, \mathcal{R}^+\}$. b is referred to as a *subject* of S iff $\forall a \in S \setminus b$, a is relevant to b.

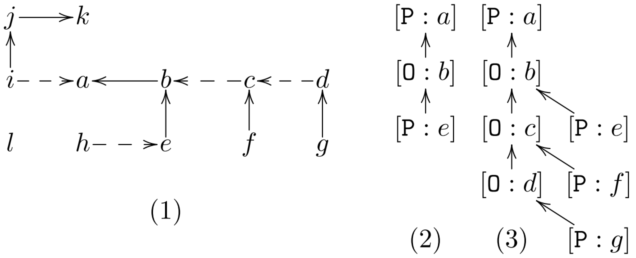

Figure 1: A BAF example adapted from (Cayrol and Lagasquie-Schiex 2005). In (1), $\longrightarrow$ represents attack relations in $\mathcal{R}^-$, $\dashrightarrow$ represents support relations in $\mathcal{R}^+$. (2) shows an abstract dispute tree. (3) shows a **bipolar** dispute tree.

Combining relevance in Defnition 4 with the uni- and bipolar admissibility semantics, we define two types of explanations in BAF. An explanation for an argument a is a special u-admissible or b-admissible set that contains a and arguments relevant to a. Formally:

Definition 5. Given a BAF framework $\langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+ \rangle$, let $S \subseteq \mathcal{A}, a \in \mathcal{A}$. Then, S is a

- *Concise Explanation (CE)* for a iff a is the subject of S, S is u-admissible, but not b-admissible;
- *Strong Explanation (SE)* for a iff a is the subject of S and S is b-admissible.

CEs are compact and shorter explanations as they only contain sufficient arguments to guard the subject and its defenders from direct attacks. SEs are stronger in a sense that they contain arguments to guard the subject and its defenders not only from all direct attacks, but also supported attacks.

Proposition 2. *Given a BAF framework, for any $a \in \mathcal{A}$ and explanation $S \subseteq \mathcal{A}$ for a, if S is a SE for a, then there exists a set $S' \subseteq S$ such that S' is a CE for a.*

Example 3. (cont'd) Consider argument a as the subject. Since j, k, l are not relevant to a, they are not included in any CEs or SEs of a. $\{a, e\}$, $\{a, e, g, h, i\}$ are CEs of a. $\{a, e, f, g\}$, $\{a, e, f, g, h, i\}$ are SEs of a.

Computing Explanations

The construction of an abstract dispute tree is an incremental process, in which the proponent attempts to counter-attack every possible attack that may come from the opponent. Hence, dispute trees can be viewed as visualization of the argumentation processes and provide useful structures from which explanations can be extracted. *Concise explanations (CEs)* can be generated from the partial framework $\langle \mathcal{A}, \mathcal{R}^- \rangle$ using abstract dispute trees as described in (Fan and Toni 2015). Here, we propose a new variant of the abstract dispute tree for computing *strong explanations (SEs)*.

Definition 6. Given a BAF framework $\langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+ \rangle$, a *bipolar dispute tree* for $a \in \mathcal{A}$ is a tree $\mathcal{T}$, such that :

1. every node of $\mathcal{T}$ is of the form $[\mathtt{L} : x]$, with $\mathtt{L} \in \{\mathtt{P}, \mathtt{O}\}$ and $x \in \mathcal{A}$: the node is *labelled* by argument x and assigned the status of either *proponent* (P) or *opponent* (O);
2. the root of $\mathcal{T}$ is a proponent node labelled by a;
3. for every proponent node $n = [\mathtt{P} : b]$, and $\forall c \in \mathcal{A}$ that $c\mathcal{R}^-b$, there exists an opponent child of n labelled by c;
4. for every opponent node $n = [\mathtt{O} : d]$: **(1)** $\forall e \in \mathcal{A}$ that $e\mathcal{R}^+d$, there exists an opponent child of n labelled by e; **(2)** there exists at most one proponent child of n, labelled by an argument f and $f\mathcal{R}^-d$;
5. no other nodes in $\mathcal{T}$ except those described in 1-4.

The set of all arguments labelling P nodes in $\mathcal{T}$ is called the *bipolar defence set* of $\mathcal{T}$, denoted by $\mathcal{D}(\mathcal{T})$.

In a bipolar dispute tree, a proponent node can only have opponent children, but an opponent node n can have at most one proponent child (attack n) and many opponent children (support n). This enables a bipolar dispute tree to represent supported attacks and "strongly defend" relationships.

Definition 7. A bipolar dispute tree $\mathcal{T}$ is a *b-admissible dispute tree* iff: **(1)** every O node in $\mathcal{T}$ has a proponent child, **(2)** no argument in $\mathcal{T}$ labels both P and O nodes, and **(3)** the closure of $\mathcal{D}(\mathcal{T})$ for $\mathcal{R}^+$ is conflict-free+. We use $\mathcal{T}_b$ to denote a b-admissible bipolar dispute tree.

Theorem 3. *Given a $BAF = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+ \rangle$ and $x \in \mathcal{A}$, let $\mathcal{T}$ be a bipolar dispute tree for x constructed from BAF.*

1. *If $\mathcal{T}$ is a b-admissible bipolar dispute tree, then $\mathcal{D}(\mathcal{T})$ is b-admissible and $\mathcal{D}(\mathcal{T})$ is a SE for x.*
2. *If $S \subseteq \mathcal{A}$ is a SE for x, then there is a b-admissible bipolar dispute tree $\mathcal{T}_b$ with its root labelled by x such that $S' = \mathcal{D}(\mathcal{T}_b)$ and $S' \subseteq S$, S' is b-admissible.*

Example 4. (cont'd) In Fig. 1, the CE computed from the abstract dispute tree in (2) is $\{a, e\}$. The SE computed from the bipolar dispute tree in (3) is $\{a, e, f, g\}$.

Conclusion

We formalized two types of explanations for acceptable arguments in Abstract Bipolar Argumentation in terms of relevance, unipolar and bipolar admissibility. *Concise explanations* are compact and tenable explanations which focus on defending direct attacks. *Strong explanations* are all-round explanations with stronger justification by defending also supported attacks. We then presented a complete and sound computational process for constructing explanations using *bipolar dispute trees*.

Acknowledgments

This research is supported, in part, by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative, the Singapore Ministry of Health under its National Innovation Challenge on Active and Confident Ageing (NIC Project No. MOH/NIC/COG04/2017 and MOH/NIC/HAIG03/2017), and the Interdisciplinary Graduate School, Nanyang Technological University under its Graduate Research Scholarship.

References

Cayrol, C., and Lagasquie-Schiex, M.-C. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *Proc. of ECSQARU (LNAI 3571)*, 378–389. Springer.

Fan, X., and Toni, F. 2015. On computing explanations in argumentation. In *Proc. of AAAI*, 1496–1502.

Zeng, Z.; Fan, X.; Miao, C.; Leung, C.; Chin, J. J.; and Ong, Y. S. 2018. Context-based and explainable decision making with argumentation. In *Proc. of AAMAS*, 1114–1122. IFAAMAS.