Argumentation Logic*

Antonis Kakas

Department of Computer Science, University of Cyprus, Cyprus
antonis@ucy.ac.cy

Argumentation Logic is born out of the growing pressure in Artificial Intelligence (AI) to develop human-like systems which have a symbiotic relationship with their users. Personal or Cognitive Assistants are required to operate typically with a Natural Language interface and to possess cognitive or thinking faculties, such as comprehension, explanation and learning that are common in the natural intelligence of people. Such systems presuppose that we are able to sufficiently formalize the human form of common sense reasoning and decision making into a logical system.

Motivated by work in Cognitive Science one approach to develop a framework for this type of informal logical reasoning is to base this on argumentation. In its most abstract form an argumentation framework in AI is defined as a tuple $< Arg, Att >$ where, Arg, is a set of arguments and Att is a binary (partial) relation on Arg, called the *attacking relation* on Arg. The attack relation is lifted in the canonical way onto the subsets of arguments, i.e. for any two subsets of arguments, A, Δ, A attacks Δ iff there exists $a \in A$ and $b \in \Delta$ such that $(a, b) \in Att$.

The central semantical notion of argumentation, namely that of a valid or acceptable argument, is given by formally capturing the statement: **"An argument is acceptable iff it renders all its attacking arguments (i.e. its counter-arguments) not acceptable"**. To do so we consider the following recursive operator of acceptability:

Definition 1. *[Acceptability Operator]*
Let $AF =< Arg, Att >$ be an abstract argumentation framework and $\mathcal{R}$ the set of binary relations on 2^{Arg}. Then the **acceptability operator**, *$\mathcal{F} : \mathcal{R} \rightarrow \mathcal{R}$, is defined as follows. For any $acc \in \mathcal{R}$ and $\Delta, \Delta_0 \in 2^{Arg}$:*

$\mathcal{F}(acc)(\Delta, \Delta_0)$ *iff*

- $\Delta \subseteq \Delta_0$, *OR,*
- *For any A such that A attacks Δ,*
  - $A \not\subseteq \Delta_0 \cup \Delta$, *AND*
  - *there exists D that attacks A such that $acc(D, \Delta_0 \cup \Delta \cup A)$.*

This operator is monotonic w.r.t. set inclusion and hence its repeated application starting from the empty binary relation has a least fixed point.

* Most of the work in this short paper comes from the articles [1, 2] and helpful discussions with Vassilis Gregoriades.

Definition 2. *[Acceptability Semantics]*
The acceptability semantics of an abstract argumentation framework AF is given by the set of **acceptable** *subsets of arguments, namely sets, Δ, such that $Acc(\Delta, \{\})$ holds where, Acc, is the least fixed point of the operator, $\mathcal{F}$, of definition 1. Acc is called the* **acceptability relation** *of AF.*

Note that the definition acceptability requires an acceptable argument Δ to (acceptably) counter-attack any counter-argument A, not only minimal sets of arguments that attack Δ. This is because the argumentation framework may contain non-deterministic choices, e.g. in the simplest case when we have two arguments a_1, a_2 which attack each other, and for an argument to be acceptable a choice between these may be needed even if neither of a_1, a_2 attack directly Δ.

Realizations of abstract argumentation, often called **structured argumentation frameworks**, assume that we are given a language, $\mathcal{L}$, for expressing information such as the premises and claims of arguments, and are required to construct the set of arguments and the attack relation between them (or directly on the subsets of arguments). In practice, we separate the attack relation into a symmetric **conflict or counter-argument** relation and a **defense** relation between arguments. The defense relation is a subset of the attack relation capturing the idea of strength between arguments, namely if "b defends against a" then the argument "b" is deemed sufficiently strong to counter back an attack coming from the argument "a". Such realizations constitute frameworks for argumentation based logics, which we call argumentation logics.

Definition 3. *[Argumentation Logic (AL) Framework]*
An ***argumentation logic framework*** *is a triple, $ALF =< Arg, \mathcal{A}, \mathcal{D} >$*[1]*:*

- *Arg is a set (of arguments)*
- *$\mathcal{A}$ is a binary (partial) relation on Arg (attack relation)*
- *$\mathcal{D}$ is a binary (partial) relation on Arg (defence relation)*

Given $A, \Delta, D \subseteq Arg$, we say that A **attacks** Δ iff there exists $a \in A$ and $b \in \Delta$ such that $(a, b) \in \mathcal{A}$ and that D **defends against** A iff $(d, c) \in \mathcal{D}$ for some $d \in D$ and $c \in A$. In addition, the empty set defends against any A iff A attacks itself, i.e. such self-attacking sets of arguments are trivially defended against and so cannot affect the semantics of the framework.

The **canonical correspondence** between an abstract argumentation framework, $AF =< Arg, Att >$, and an argumentation logic realization, $ALF =< Arg, \mathcal{A}, \mathcal{D} >$ is given by minimally applying the rules:

$Att(a, b) \Rightarrow (a, b) \in \mathcal{A}$, $(b, a) \in \mathcal{A}$ and $(a, b) \in \mathcal{D}$

$(a, b) \in \mathcal{A}$ and $((b, a) \in \mathcal{D} \rightarrow (a, b) \in \mathcal{D}) \Rightarrow Att(a, b)$.

Definition 4. *[Dialectic Acceptability Operator for AL]*
Let $ALF =< Arg, \mathcal{A}, \mathcal{D} >$ be an argumentation logic framework and $\mathcal{R}$ the set of binary relations on 2^{Arg}. Then the **L-acceptability operator***, $\mathcal{LF} : \mathcal{R} \rightarrow \mathcal{R}$,*

[1] Normally, $\mathcal{D} \subseteq \mathcal{A}$.

is defined as follows. For any $acc \in \mathcal{R}$ *and* $\Delta, \Delta_0 \in 2^{Arg}$*:*

$(\Delta, \Delta_0) \in \mathcal{LF}(acc)$ *iff*

- $\Delta \subseteq \Delta_0$*, or,*
- *For any* A *such that* A *attacks* Δ*,*
  - $A \not\subseteq \Delta_0 \cup \Delta$*, and*
  - *there exists* D *such that* D *defends against* A *and* $(D, \Delta_0 \cup \Delta) \in acc$*.*

This is monotonic operator and its least fixed point, $LAcc$ gives the semantics of the argumentation logic framework, ALF. A set Δ of arguments is **logically acceptable** or simply **L-acceptable** iff $LAcc(\Delta, \{\})$ holds. **Logical entailment** of an argument a in ALF is then defined by requiring that there is an L-acceptable set Δ in which the argument belongs (a **is credulously entailed**) and for any conflicting argument a^c, i.e. such that $(a^c, a) \in \mathcal{A}$, there is no acceptable set in which a^c belongs (**sceptically entailed**).

The following "propositional" case(s) of an ALF gives Argumentation Logic(s).

Definition 5. *[Propositional Argumentation Logic(s)]*
Let L *be the language of Propositional Logic and* $PAL =< Arg, \mathcal{A}, \mathcal{D} >$ *be the argumentation logic framework, called the* ***Propositional Argumentation Logic****, given as follows:*

- *Arg is the set of propositional formulae*
- $(A, B) \in \mathcal{A}$ *iff* $A \cup B \vdash_{L0} \perp$
- $(D, A) \in \mathcal{D}$ *iff*
  - $D = \{\neg\phi\}$ *(resp.* $D = \{\phi\}$*) and* $\phi \in A$ *(resp.* $\neg\phi \in A$*), or*
  - $D = \{\}$ *and* $A \vdash_{L0} \perp$

where $\vdash_{L0}$ *is a chosen sub-logic of Propositional Logic, which we call the* ***core logic*** *of* PAL*, and denote by* $PAL(\vdash_{L0})$*.*

The canonical choice for $\vdash_{L0}$ is that of **direct derivation,** $\vdash_{DD}$, i.e. given by the Natural Deduction proof rules minus the negation introduction $\neg I$ proof rule for negation introduction, i.e. the Reductio ad Absurdum proof rule. Logical entailment of a propositional formula ϕ under PAL is then given simply by requiring that $\{\phi\}$ is acceptable in PAL and that $\{\neg\phi\}$ is not acceptable in PAL. We can then show that PAL can capture classical Propositional Logic.

Theorem 1. $PAL(\vdash_{DD})$ *is logically equivalent to classical Propositional Logic.*

When a given propositional theory is inconsistent, PAL deviates from Propositional Logic, as like any argumentation logic, PAL does not trivialize under conflicting premise information.

There are two essential questions that we want to ask about argumentation logic frameworks, $ALF =< Arg, \mathcal{A}, \mathcal{D} >$.

(Q1) Under what conditions on the attack and defence relations in ALF the L-acceptability of a set of arguments in ALF implies the acceptability of the same set (or a superset of this) in the corresponding abstract argumentation framework, AF. In other words, when is ALF a sound approximation of the corresponding AF.

To study this we first note that for L-acceptability we can restrict attention only at minimal attacks. The non-deterministic choice is transferred to the choice of defenses in the L-acceptability of a set of arguments. The above question then can be posed as the question of when the global union, $M = \bigcup_i D_i$, of all the defences in the L-acceptability of a set of arguments Δ is conflict free (does not attack itself). This is the main **rationality or coherence property** required of an argumentation logic framework. When this holds then L-acceptability implies acceptability and these global sets of all defenses form the **argumentation models** of the argumentation logic framework.

We can show that when an $ALF =< Arg, \mathcal{A}, \mathcal{D} >$ satisfies the property that for any argument, a, there exists a unique argument, a^c such that (a, a^c) or $(a^c, a) \in \mathcal{D}$ then the set of defences of any L-acceptable set of arguments Δ is conflict free and there exists a superset of Δ which is acceptable in the corresponding abstract acceptability framework. A special case of this result holds for propositional argumentation logic frameworks and indeed the argumentation models of PAL coincide with the two valued propositional models for any given propositional theory of premises T which is classically consistent.

(Q2) How incomplete is the $ALF =< Arg, \mathcal{A}, \mathcal{D} >$ framework with respect to its corresponding abstract argumentation framework? To illustrate this incompleteness consider a simple example of an abstract argumentation framework $AF =< Arg, Att >$ where $Arg = \{a, b, b^c, c\}$ and $Att = \{(a, c), (b, a)(b^c, b), (a, b^c)\}$. The corresponding $ALF =< Arg, \mathcal{A}, \mathcal{D} >$ has

$\mathcal{A} = \{(a, c), (b, a)(b^c, b), (a, b^c), (c, a), (a, b)(b, b^c), (b^c, a)\}$

$\mathcal{D} = \{(a, c), (b, a)(b^c, b), (a, b^c)\}$.

In $AF =< Arg, Att >$ the set $\{c\}$ is acceptable, i.e $Acc(\{c\}, \{\})$ holds: its only attack by $\{a\}$ is counter-attacked by $\{b\}$ which in turn is acceptable relative to $\{a\}$, i.e. $Acc(\{b\}, \{a\})$ holds since any attack against $\{b\}$, must contain $\{b^c\}$ which is counter-attacked by $\{a\}$. Note also that any superset of the attack $\{a\}$ is self-attacking so it is trivially defended). But in the corresponding $ALF =< Arg, \mathcal{A}, \mathcal{D} >$ framework $LAcc(\{c\}, \{\})$ does not hold. The attack $\{a\}$ can be defended by $\{b\}$ but its attack by $\{b^c\}$ can only be defended by $\{a\}$ and $LAcc(\{a\}, \{c, b\})$ does not hold. This incompleteness comes from the fact that L-acceptability cannot recognize the non acceptability of the argument $\{a\}$ that comes from the fact that this is an internally self-defeating fallacious argument and hence there is no need to find an external explicit defense against it.

References

1. A. Kakas and P. Mancarella. On the Semantics of Abstract Argumentation. *Journal of Logic and Computation*, 23(5):991–1015, 2013.
2. Antonis C. Kakas, Paolo Mancarella, and Francesca Toni. On argumentation logic and propositional logic. *Studia Logica*, 106(2):237–279, 2018.