On Acceptability in Abstract Argumentation Frameworks with an Extended Defeat Relation

Diego C. Martínez Alejandro J. García Guillermo R. Simari
Artificial Intelligence Research and Development Laboratory,
Department of Computer Science and Engineering, Universidad Nacional del Sur,
Email: {dcm,ajg,grs}@cs.uns.edu.ar

Abstract. Defeat between arguments is established by a combination of two basic elements: a conflict or defeat relation, and a preference relation on the arguments involved in this conflict. We present a new abstract framework for argumentation where two kinds of defeat are present, depending on the outcome of the preference relation: an argument may be a *proper defeater* or a *blocking defeater* of another argument. An operator is used to characterize the set of accepted arguments. This operator also provides a method for identifying controversial situations.

Keywords. Abstract argumentation, argumentation semantic, preference relation.

1. Introduction

The area of Knowledge Representation and Reasoning has been enriched during the past two decades with the addition of Argument-Based Reasoning Systems [1,2,3] to mention a few. Two interesting surveys on argumentation are [4,5] and the reader is referred to them for details on the different proposals.

The study of the acceptability of arguments is one of the main concerns in Argumentation Theory. In formal systems of defeasible argumentation, arguments for and against a proposition are produced and evaluated to test the acceptability of that proposition following a dialectical process [6]. The main idea in these systems is that a proposition will be accepted as true if there exists an argument that supports it, and this argument is acceptable according to an analysis between it and its counterarguments. This analysis requires a process of comparison of conflicting arguments in order to decide which one is preferable [1,7,8,9,10]. After this dialectical analysis is performed over the set of arguments in the system, some of them will be *acceptable* arguments, while others will be not. Argumentation is used as a form of non-monotonic or defeasible reasoning [11] and it is suitable for modeling dialogues between intelligent agents [12].

Abstract argumentation systems [13,3,9] are formalisms for argumentation where some components remain unspecified. Usually, the actual structure of an argument is abstracted away. In this kind of system, the emphasis is put on the semantic notion of finding the set of accepted arguments. Most of them are based on the single abstract concept of the *attack* represented as a binary relation, and extensions are defined as sets

of possibly accepted arguments. This primitive notion of defeat between arguments is the basis of the study of argumentation semantic, but a more detailed model will be useful to capture specific behaviour of concrete systems.

We define a framework where the defeat relation between arguments is decomposed into two basic elements: symmetric conflicts and a preference criterion. Finding a preferred argument is essential to determine a defeat relation [1,8,9,10]. However, the task of comparing arguments to establish a preference is not always successful. In this case, the classic abstract attack relation is no longer useful as a modelling tool. In the next section, we present an abstract framework for argumentation where conflicts and preference between arguments are considered, and the associated semantic operator is defined.

2. Argumentation Framework

Our argumentation framework is formed by four elements: a set of arguments, and three basic relations between arguments.

Definition 1 *An abstract argumentation framework (AF) is a quartet* $\langle Args, \sqsubseteq, \mathbf{C}, \mathbf{R}\rangle$, *where* $Args$ *is a finite set of arguments,* $\sqsubseteq$ *is the subargument relation,* $\mathbf{C}$ *is a symmetric and anti-reflexive binary conflict relation between arguments,* $\mathbf{C} \subseteq Args \times Args$, *and* $\mathbf{R}$ *is a preference relation among arguments.*

Here, arguments are abstract entities [13] that will be denoted using calligraphic uppercase letters. No reference to the underlying logic is needed since we are abstracting the structure of the arguments (see [1,8,11] for concrete systems). The symbol $\sqsubseteq$ denotes subargument relation: $\mathcal{A} \sqsubseteq \mathcal{B}$ means "$\mathcal{A}$ is a subargument of $\mathcal{B}$".

The conflict relation between two arguments $\mathcal{A}$ and $\mathcal{B}$ denotes the fact that these arguments cannot be accepted simultaneously since they contradict each other. For example, two arguments $\mathcal{A}$ and $\mathcal{B}$ that support complementary conclusions l and $\neg l$ cannot be accepted together. Also an argument with hypotesis h cannot be accepted together with an argument for $\neg h$. The set of all pairs of arguments in conflict on Φ is denoted by $\mathbf{C}$. Given a set of arguments S, an argument $\mathcal{A} \in S$ is said to be in conflict in S if there is an argument $\mathcal{B} \in S$ such that $(\mathcal{A}, \mathcal{B}) \in \mathbf{C}$. The set *Conf*$(\mathcal{A})$ is the set of all arguments $\mathcal{X} \in Args$ such that $(\mathcal{A}, \mathcal{X}) \in \mathbf{C}$.

The constraints imposed by the conflict relation lead to several sets of possible accepted arguments. Therefore, some way of deciding among all the possible outcomes must be devised. In order to accomplish this task, the relation $\mathbf{R}$ is introduced in the framework and it is used to evaluate arguments, modelling a preference criterion based on a measure of strength. If $\mathcal{A}\mathbf{R}\mathcal{B}$ but not $\mathcal{B}\mathbf{R}\mathcal{A}$ then $\mathcal{A}$ is preferred to $\mathcal{B}$, denoted $\mathcal{A} \succ \mathcal{B}$. If $\mathcal{A}\mathbf{R}\mathcal{B}$ and $\mathcal{B}\mathbf{R}\mathcal{A}$ then $\mathcal{A}$ and $\mathcal{B}$ are arguments with equal relative preference, denoted $\mathcal{A} \equiv \mathcal{B}$. If neither $\mathcal{A}\mathbf{R}\mathcal{B}$ or $\mathcal{B}\mathbf{R}\mathcal{A}$ then $\mathcal{A}$ and $\mathcal{B}$ are incomparable, denoted $\mathcal{A} \bowtie \mathcal{B}$.

Preference is usually based on structural properties of arguments, as the number of logical rules used to derive the conclusion or the number of propositions involved in that process. Other non-trivial preferences may be captured by $\mathbf{R}$, for example, the fact that an argument with conclusion $\neg h$ is preferred to an argument whith hypotesis h. As the comparison criterion is treated abstractly, we do not assume any property of $\mathbf{R}$ but, as stated in [3], several conditions must be satisfied, for example, that an argument is always preferred (or equivalent in conclusive force) to any superargument. Therefore, if $\mathcal{A} \succ \mathcal{B}$ then $\mathcal{A} \succ \mathcal{C}$ for any superargument $\mathcal{C}$ of $\mathcal{B}$. Any concrete framework may

establish additional requirements for decision making. The conflict relation should also exhibit a rational behaviour regarding subarguments. If $(\mathcal{A}, \mathcal{B}) \in \mathbf{C}$, then $(\mathcal{A}, \mathcal{B}_1) \in \mathbf{C}$, $(\mathcal{A}_1, \mathcal{B}) \in \mathbf{C}$, and $(\mathcal{A}_1, \mathcal{B}_1) \in \mathbf{C}$, for any arguments $\mathcal{A}_1, \mathcal{B}_1$, $\mathcal{A} \sqsubseteq \mathcal{A}_1$ and $\mathcal{B} \sqsubseteq \mathcal{B}_1$. We call this property *conflict inheritance*: if an argument $\mathcal{A}$ is in conflict with an argument $\mathcal{B}$ then that conflict is still present when considering superarguments of $\mathcal{A}$ or $\mathcal{B}$.

Example 1 *$\Phi = \langle Args, \sqsubseteq, \mathbf{C}, \mathbf{R} \rangle$ is an AF where $Args = \{\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}\}$, $\mathbf{C} = \{\{\mathcal{A}, \mathcal{B}\}, \{\mathcal{B}, \mathcal{C}\}, \{\mathcal{C}, \mathcal{D}\}\}, \{\mathcal{C}, \mathcal{E}\}\}$[1] and $\mathcal{A} \succ \mathcal{B}, \mathcal{B} \succ \mathcal{C}$, $\mathcal{E} \bowtie \mathcal{C}$ and $\mathcal{C} \equiv \mathcal{D}$.*

For two arguments $\mathcal{A}$ and $\mathcal{B}$ in $Args$, such that the pair $(\mathcal{A}, \mathcal{B})$ belongs to $\mathbf{C}$ the relation $\mathbf{R}$ is considered. If a concrete preference is made ($\mathcal{A} \succ \mathcal{B}$ or $\mathcal{B} \succ \mathcal{A}$), then a defeat relation is established. It is said that the preferred argument is a *proper defeater* of the non-preferred argument. If the arguments are *indifferent* according to $\mathbf{R}$, then they have the *same* relative conclusive force. For example, if the preference criterion establishes that smaller arguments are preferred, then two arguments of the same size are indifferent. On the other hand, arguments may be *incomparable*. For example, if the preference criterion states that argument $\mathcal{A}$ is preferred to $\mathcal{B}$ whenever the premises of $\mathcal{A}$ are included in the premises of $\mathcal{B}$, then arguments with disjoint sets of premises are incomparable. This situation must be understood as a natural behaviour. When two conflictive arguments are indifferent or incomparable according to $\mathbf{R}$, the conflict between these two arguments remains unresolved. Due to this situation and to the fact that the conflict relation is a symmetric relation, each of the arguments is *blocking* the other one and it is said that both of them are *blocking defeaters* [1]. An argument $\mathcal{B}$ is said to be a *defeater* of an argument $\mathcal{A}$ if $\mathcal{B}$ is a blocking or a proper defeater of $\mathcal{A}$. In example 1, argument $\mathcal{A}$ is a proper defeater of argument $\mathcal{B}$, while $\mathcal{C}$ is a blocking defeater of $\mathcal{D}$ and vice versa.

Well known semantics for abstract argumentation frameworks are based on defeat relations, usually called *attack* relations [13,3,14]. These formalisms assume the existence of a binary relation of attack (not necessarily symmetric) defined over the set of all possible arguments, such that if $(\mathcal{A}, \mathcal{B})$ are in the attack relation then in order to accept $\mathcal{B}$ it is necessary to find out if $\mathcal{A}$ is accepted or not, but not the other way around. The acceptance relation should be derived from a conflict relation between arguments and a suitable comparison criterion, and that criterion usually remains unspecified in the abstract system. This remark on the attack relation is seldom made. It is our contention that an extended semantics for argumentation will be useful. This semantics will be based on the two defining characteristics of an argumentation system: the conflict relation between arguments and the comparison criterion used to evaluate such arguments.

Arguments can be classified as *accepted* arguments or *non-accepted* or *rejected* arguments according to their context in the framework. Any set of accepted arguments should not contain arguments in conflict. A set of arguments S is said to be *conflict free* if for all $\mathcal{A}, \mathcal{B} \in S$ then $(\mathcal{A}, \mathcal{B}) \notin \mathbf{C}$. In example 1 the set $\{\mathcal{A}, \mathcal{C}\}$ is a conflict free set.

Given a set of arguments S, two kinds of arguments are easily identified as accepted arguments: first, those arguments not involved in any conflict in S; second, those arguments actually involved in a conflict, but preferred to the arguments that are in conflict with them, according to relation $\mathbf{R}$. Both kinds of special arguments are called *defeater free* arguments. An argument $\mathcal{A}$ is defeater-free in a set S if no argument in S is a de-

[1] When describing elements of $\mathbf{C}$, we write $\{\mathcal{A}, \mathcal{B}\}$ as an abbreviation for $\{(\mathcal{A}, \mathcal{B}), (\mathcal{B}, \mathcal{A})\}$, for any arguments $\mathcal{A}$ and $\mathcal{B}$ in $Args$.

feater of $\mathcal{A}$. Defeater-free arguments must be accepted, since no (preferred) contradictory information is provided in the framework. Note that this classification is relative to the set in which the argument is included. The semantic of $\mathbf{C}$ states that when an argument $\mathcal{A}$ is accepted, any argument in *Conf*($\mathcal{A}$) should be rejected. The following definition captures a subset of arguments that should be rejected in the framework.

Definition 2 *Let S be a set of arguments in $\langle Args, \sqsubseteq, \mathbf{C}, \mathbf{R}\rangle$. An argument $\mathcal{A} \in S$ is said to be suppressed in S if one of the following cases hold: (a) there is a defeater-free argument $\mathcal{B}$ in S such that $\mathcal{B}$ is a proper defeater of $\mathcal{A}$, or (b) there is a blocking defeater $\mathcal{B}$ of $\mathcal{A}$in S, and there is no other argument $\mathcal{C}$ ($\mathcal{C} \neq \mathcal{A}$) in S such that $\mathcal{C}$ is a defeater of $\mathcal{B}$.*

The first case is clear since any argument involved in a conflict must be suppressed when its counterpart in this conflict is accepted (has no defeater). The second case reflects the situation in which two arguments are taking part of an unsolved conflict and from the point of view of one of them ($\mathcal{A}$) its opponent is not attacked by a third argument. The argument $\mathcal{A}$ should be suppressed since the threat of $\mathcal{B}$ cannot be avoided, despite other attacks on $\mathcal{A}$. Note that if $\mathcal{A}$ is only defeated by $\mathcal{B}$ then both arguments should be suppressed because the blocking condition is symmetrical.

Given a set S of arguments it is as easy to identify obviously suppressed arguments as it is to identify inevitably accepted ones. The following function $\Upsilon : 2^{Args} \longrightarrow 2^{Args}$ characterizes the set of arguments not directly suppressed in a given set S.

$$\Upsilon(S) = \{\mathcal{A} : \mathcal{A} \in S \text{ and } \mathcal{A} \text{ is not suppressed in } S\}$$

It is easy to see that if S is a conflict-free set of arguments, then $S = \Upsilon(S)$. However, the converse is not true, as shown in the next example:

Example 2 *Let $\langle Args, \sqsubseteq, \mathbf{C}, \mathbf{R}\rangle$ be an AF, where $Args = \{\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}\}$ and $\mathbf{C} = \{\{\mathcal{A}, \mathcal{B}\}, \{\mathcal{B}, \mathcal{C}\}, \{\mathcal{C}, \mathcal{D}\}, \{\mathcal{D}, \mathcal{A}\}\}$ and for all arguments $\mathcal{X}$ and $\mathcal{Y}$, $\mathcal{X} \bowtie \mathcal{Y}$. No argument in $Args$ is a defeater-free argument, therefore $\Upsilon(Args) = Args$.*

By definition, $\Upsilon(S)$ includes some (or all) of the arguments in S. In the set $\Upsilon(S)$ some arguments may now be classified as *defeater-free* arguments, since its defeaters are suppressed arguments in S. It is then possible to repeatedly apply function Υ to the set of arguments in the framework. This process may continue until a fixpoint is reached.

Definition 3 *Υ^n is defined as: Υ^0 is $Args$, and $\Upsilon^{(n+1)} = \Upsilon \circ \Upsilon^n$. The set of arguments $\Upsilon^k, k \geq 0$ such that $\Upsilon^k = \Upsilon^{k+1}$ is denoted Υ^ω.*

Example 3 *Let $\Phi_2 = \langle Args, \sqsubseteq, \mathbf{C}, \mathbf{R}\rangle$ be an AF where $Args = \{\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}\}$, $\mathbf{C} = \{\{\mathcal{A}, \mathcal{B}\}, \{\mathcal{B}, \mathcal{C}\}, \{\mathcal{C}, \mathcal{D}\}\}$ and $\mathcal{A} \equiv \mathcal{B}$, $\mathcal{B} \bowtie \mathcal{C}$ and $\mathcal{C} \succ \mathcal{D}$. In this framework, $\Upsilon^1 = \{\mathcal{A}, \mathcal{D}, \mathcal{C}\}$, because $\mathcal{B}$ is a suppressed argument, as $\mathcal{A}$ is a blocking defeater not defeated by a third argument. $\Upsilon^2 = \{\mathcal{A}, \mathcal{C}\}$ because $\mathcal{D}$ is defeated by $\mathcal{C}$ which is now defeater-free in Υ^1. Because $\Upsilon^2 = \Upsilon^3$ then $\Upsilon^\omega = \{\mathcal{A}, \mathcal{C}\}$.*

Trivially, no argument is suppressed in Υ^ω. An argument in Υ^ω which is not in conflict with any other argument in the same set is an accepted argument. The set of accepted arguments in Υ^ω is denoted $\Upsilon^{\omega+}$. Therefore, if Υ^ω is a conflict-free set (as in example 3, but not in example 2), then any argument in Υ^ω is an *accepted* argument.

The previously defined conflict inheritance leads to a common sense property of argumentation frameworks. For any argument $\mathcal{A}$, if $\mathcal{A} \in \Upsilon^{\omega+}$ then $\mathcal{B} \in \Upsilon^{\omega+}$ for all $\mathcal{B} \sqsubseteq \mathcal{A}$. Suppose $\mathcal{A}_1 \sqsubseteq \mathcal{A}$ is not in Υ^{ω}. Then $\mathcal{A}_1$ is a suppressed argument, because one of the conditions of definition 2 holds in some $\Upsilon^i, i > 0$. But if $\mathcal{A}_1$ is suppressed in Υ^i then also $\mathcal{A}$ is suppressed in Υ^i because they share defeaters (because of conflict inheritance) and therefore is also suppressed. The reader is referred to [15] for the role of subarguments in well structured argumentation, using the framework of definition 1.

In the framework of example 2, no arguments should be accepted as it is not possible to establish a concrete preference. Here, Υ^{ω} is not a conflict-free set. This is related to the presence of some special arguments involved in a cicle of defeaters, a common situation called a *fallacy*. Any argument involved in a fallacy is usually called *fallacious*. The most important premise in defeasible argumentation is that an argument must be *accepted* only when none of its defeaters are. However, no fallacious argument can exhibit this property, because at least one of its defeaters is also a fallacious argument [2]. Therefore, any argument of this kind should not be accepted. An AF is said to contain a fallacy if Υ^{ω} is not a conflict-free set of arguments.

3. Related Work and Conclusions[3]

We introduced a new abstract framework for argumentation where two kinds of defeat are present, depending on the outcome of the preference relation. A fix-point operator is used to characterize the set of accepted arguments. This operator also provides a method for identifying controversial situations.

Since the introduction of Dung's seminal work [13] on the semantics of argumentation this area has been extremely active. This approach begins by defining an abstract framework in order to characterize the set of accepted arguments independently of the underlying logic. We followed this line in this work. In Dung's presentation no explicit preference relation is included, and the basic interaction between arguments is the binary, non-symmetric, *attack* relation. This style of argument attack is used in a number of different abstract frameworks, but none of them separates the notion of preference criteria from the conflict relation, as it is usually done in concrete systems. The classic attack relation allows the definition of mutual defeaters: two arguments attacking each other. This is not very realistic, as there is not an attack situation (in the sense of being conflictive and preferred to the opponent) but a controversial situation due to the lack of decision in the system. In our framework, this leads to blocking defeaters. The fixpoint semantic defined here results more credulous than the classic grounded extension [13], as it can be noted in example 3, where according to Dung the grounded extension is the empty set.

Several frameworks do include a preference relation. Vreeswijk, in [3], defines a popular abstract framework, making important considerations on comparison criterions. Interesting frameworks that consider the issue of preference relations are introduced in [9], [16] and in [17]. In these frameworks the basic interaction between agents is the classic *attack* relation, and the preference order is used as a defense against conflictive arguments. The defeat relation arises when the preferences agree with the attack.

[2] Because any non-fallacious defeater has been previously suppressed.

[3] Space limitations prevent us of a more complete review of related work

Bench-Capon, in [18], also defines an argumentation framework that includes a way to compare arguments. A set of values related to arguments is defined in the framework. Since a preference relation is defined on the values promoted by arguments, those arguments can be weighted in order to resolve attacks. However, only a single notion of defeat is derived. This defeat occurs when the value promoted by the attacked argument is not preferred to the value promoted by the attacker. Again, the preference order is used to check if the attacker argument is preferred, not to elucidate symmetric conflicts as it is used in our framework.

References

[1] Guillermo R. Simari and Ronald P. Loui. A Mathematical Treatment of Defeasible Reasoning and its Implementation. *Artificial Intelligence*, 53:125–157, 1992.

[2] Robert A. Kowalski and Francesca Toni. Abstract argumentation. *Artificial Intelligence and Law*, 4(3-4):275–296, 1996.

[3] Gerard A. W. Vreeswijk. Abstract argumentation systems. *Artificial Intelligence*, 90(1–2):225–279, 1997.

[4] Carlos I. Chesñevar, Ana G. Maguitman, and Ronald P. Loui. Logical Models of Argument. *ACM Computing Surveys*, 32(4):337–383, December 2000.

[5] Henry Prakken and Gerard Vreeswijk. Logical systems for defeasible argumentation. In D.Gabbay, editor, *Handbook of Philosophical Logic, 2nd ed.* Kluwer Academic Pub., 2000.

[6] G.R. Simari, C.I. Chesñevar, and A.J. García. The role of dialectics in defeasible argumentation. In *XIV Intl. Conf. of the Chilenean Computer Science Society* , pages 111–121, 1994.

[7] David L. Poole. On the Comparison of Theories: Preferring the Most Specific Explanation. In *Proc. 9th IJCAI*, pages 144–147. IJCAI, 1985.

[8] Henry Prakken and Giovanni Sartor. A system for defeasible argumentation, with defeasible priorities. In *Proc. of the Intl. Conf. on Formal and Applied Practical Reasoning (FAPR-96)* , volume 1085 of *LNAI*, pages 510–524, June 3–7 1996.

[9] Leila Amgoud. Using preferences to select acceptable arguments. In *Proc. of European Conf. in Artificial Intelligence (ECAI'98) , Brighton* , pages 43–44, August 1998.

[10] F. Stolzenburg, A.J. García, C.I. Chesñevar, and G.R. Simari. Computing generalized specificity. *Journal of Aplied Non-Classical Logics* , 13(1):87–113, January 2003.

[11] Alejandro J. García and Guillermo R. Simari. Defeasible logic programming: An argumentative approach. *Theory and Practice of Logic Programming*, 4(1):95–138, 2004.

[12] Leila Amgoud, Simon Parsons, and Nicolas Maudet. Arguments, dialogue and negotiation. In *Proc. of the 14th European Conf. on Artificial Intelligence* , pages 338–342. ECAI, 2001.

[13] Phan M. Dung. On the Acceptability of Arguments and its Fundamental Role in Nomonotonic Reasoning and Logic Programming. In *Proc. of the 13th IJCAI*, pages 852–857, 1993.

[14] C. Cayrol, S. Doutre, M. C. Lagasquie-Schiex, and J. Mengin. "Minimal Defence": a refinement of the preferred semantics for argumentation frameworks. In *Proc. of the 9th Intl. Workshop on Non-Monotonic Reasoning* , pages 408–415, July 2002.

[15] D.C. Martínez, A.J. García, and G.R. Simari. Progressive defeat paths in abstract argumentation frameworks. In *Proc. of the 19th Conf. of the Canadian Society for Computational Studies of Intelligence*, pages 242–253, 2006.

[16] Leila Amgoud and Claudette Cayrol. On the acceptability of arguments in preference-based argumentation. In *14th Conf. on Uncertainty in Artificial Intelligence* , pages 1–7, 1998.

[17] Leila Amgoud and Laurent Perrussel. Arguments and Contextual Preferences. In *Computational Dialectics-Ecai workshop (CD2000) , Berlin* , August 2000.

[18] T.J.M. Bench-Capon. Value-based argumentation frameworks. In *Proc. of Nonmonotonic Reasoning*, pages 444–453, 2002.