Meta-Argumentation Frameworks for Multi-party Dialogues

Gideon Ogunniye[1], Alice Toniolo[2], and Nir Oren[1]

[1] Department of Computing Science, University of Aberdeen, Scotland, UK
[2] School of Computer Science, University of St Andrews, Scotland, UK

Abstract. The conclusions drawn from a dialogue depend both on the content of the arguments, and the level of trust placed in the arguments and the entity advancing them. In this paper, we describe a framework for dialogue where such trust forms the basis for expressing preferences between arguments, and in turn, for computing conclusions of the dialogue. Our framework contains object and meta-level arguments, and uses ASPIC+ to represent arguments, while argument schemes capture meta-level arguments about trust and preferences.

1 Introduction

In human dialogue, conclusions are drawn not only based on argument interactions, but also by considering the level of trust or confidence placed in the arguments and those presenting them. Critically, as the dialogue progresses, additional utterances can cause these levels of trust to change, and capturing such changes is therefore important.

Since we consider the arguments advanced during the dialogue, as well as argument about those arguments, our approach builds on Muller's meta-argumentation system [7]. Here, *object-level* arguments are advanced which deal with the topic of the dialogue. *Meta-level* arguments then describe arguments about arguments, including whether an argument attacks another; what counts as an argument; and whether an argument is preferred over another. Our focus in this paper involves arguments which relate to trust between arguments, and we consider several such classes of argument, described through argumentation schemes. As the dialogue progresses, arguments attacking and supporting these arguments can be introduced, causing shifts in trust over time, in contrast to systems such as [2, 3, 11], where preferences and trust in arguments are fixed.

Our work combines several existing frameworks and techniques, and in the next section, we provide the background necessary to our approach. In Section 3, we introduce our dialogue model and the argument schemes used within our meta-argumentation framework. Section 4 discusses an example of our work and we conclude in Section 5.

2 Background

Our work builds on a fragment of ASPIC+ [6], which uses abstract argumentation [4] to identify justified conclusions. We therefore begin by briefly discussing these.

Definition 1. *An* argument framework *(AF) is a pair $\langle \mathcal{A}, \mathcal{D} \rangle$ where $\mathcal{A}$ is a set of arguments and $\mathcal{D} \subseteq \mathcal{A} \times \mathcal{A}$ is a binary defeat relation. Given $AF = \langle \mathcal{A}, \mathcal{D} \rangle$, and $\mathcal{E} \subseteq \mathcal{A}$,*

- *$\mathcal{E}$ is conflict-free iff there are no $\phi_1, \phi_2 \in \mathcal{E}$ s.t. $(\phi_1, \phi_2) \in \mathcal{D}$.*
- *$\mathcal{E}$ defends ϕ_1 iff for every $(\phi_2, \phi_1) \in \mathcal{D}$, there exist a $\phi_3 \in \mathcal{E}$ s.t. $(\phi_3, \phi_2) \in \mathcal{D}$.*
- *$\mathcal{E}$ is an admissible set iff $\mathcal{E}$ is conflict free and defends all its elements.*
- *$\mathcal{E}$ is a complete extension iff there are no other elements which it defends.*
- *$\mathcal{E}$ is a preferred extension iff it is a maximal complete extension.*

An extension identifies a consistent set of arguments and conclusions. While many different classes of extensions have been defined, we focus on preferred extensions here. It should be noted that an AF can have multiple different preferred extensions. An argument present in all extensions is *sceptically* justified; if it is present in at least one extension, it is *credulously* justified.

AFs as described above are abstract and lack structure. Given a knowledge base, we must be able to determine which arguments can be constructed, and for this purpose, we make use of a fragment the popular ASPIC+ framework [6]. ASPIC+ defines an *argumentation system* built from an (unspecified) logical language $\mathcal{L}$ which is closed under negation ($\neg$). Arguments are then formed by repeatedly applying strict (elements of R_s) or defeasible (R_d) inference rules to elements from a knowledge base $\mathcal{K}$. The argumentation system contains a function $n : R_d \rightarrow \mathcal{K}$, associating defeasible rules with entities in the knowledge base. Arguments in ASPIC+ attack each other when inconsistencies exist between them. ASPIC+ describes how preferences between arguments can be obtained from preferences between rules and elements in the knowledge base determining successful attacks; i.e. defeats. The resultant structure is referred to as an *argumentation theory*, corresponding to an argumentation framework as per Definition 1. In our approach, we consider only defeasible rules, no preferences, and assume that all elements in a knowledge base can be attacked.

Definition 2. *(Argument and Attack) [6]. An argument A on the basis of a knowledge base $\mathcal{K}$ in argumentation system $(\mathcal{L}, \neg, \mathcal{R}_d, n)$ is*

1. *μ if $\mu \in \mathcal{K}$ with: $Prem(A) = \{\mu\}$, $Conc(A) = \{\mu\}$, $Sub(A) = \{\mu\}$.*
2. *$A_1, \ldots, A_n \rightarrow / \Rightarrow \psi$ if $A_1, \ldots, A_n$ are arguments such that there exists a a defeasible rule $Conc(A_1), \ldots, Conc(A_n) \Rightarrow \psi$ in $\mathcal{R}_d$ with $Prem(A) = Prem(A_1) \cup \ldots \cup Prem(A_n)$, $Conc(A) = \{\mu\}$, $Sub(A) = Sub(A_1) \cup \ldots \cup Sub(A_n) \cup \{A\}$.*
3. *A attacks B iff A undercuts, rebuts or undermines B, where A undercuts B (on B') iff $Conc(A) = \neg n(r)$ for some $B' \in Sub(B)$. A rebuts B (on B') iff $Conc(A) = \neg\mu$ for some $B' \in Sub(B)$ of the form $B''_1, \ldots, B''_n \Rightarrow \mu$. A undermines B (on μ) iff $Conc(A) = \neg\mu$ for a premise μ of B for an ordinary premise μ of B.*

3 Hierarchical Systems of Arguments and Dialogues

Our approach uses meta-level arguments about trust. These refer to object-level arguments about the original dialogue topic. We build on the ideas of Wooldridge [14], who suggested that arguments and dialogue are inherently meta-logical processes. Thus, arguments advanced in a dialogue are not restricted to asserting the truth or falsity of statements, but include arguments about arguments; taking a hierarchical view, arguments at level n of the hierarchy may refer to the same or lower levels in the hierarchy.

In our work, we consider a hierarchy with 3 levels, labelled $l_0, \ldots l_2$. The object level (l_0) contains arguments and attacks related to the domain of discourse. Arguments at level l_1 support arguments at the object level and indirectly attack them by attacking other arguments within l_1. These capture the trust placed in object level arguments and attacks. Similarly, arguments at l_2 attack others in this level, as well as at level l_1, and capture trust in sources of object-level arguments. All of these arguments and the interactions between them are encoded in a *bimodal argument graph*.

3.1 Bimodal Argument Graphs

A bimodal argument graph is a hierarchical structure capturing object and meta-level arguments, and the attacks and supports between them.

Definition 3. *A* Bimodal Argument Graph *for a reasoner Ag_I is a tuple $BAG_I = \langle \mathcal{A}_O, \mathcal{A}_{M_I}, \mathcal{D}_O, \mathcal{D}_{M_I}, \mathcal{S}_{MO_I}, \mathcal{S}_{MA_I} \rangle$ where*

- *$\mathcal{A}_O$ and $\mathcal{A}_{M_I}$ are object-level and meta-level arguments respectively such that $\mathcal{A}_O \cap \mathcal{A}_{M_I} = \emptyset$.*
- *$\mathcal{D}_O \subseteq \mathcal{A}_O \times \mathcal{A}_O$ and $\mathcal{D}_{M_I} \subseteq \mathcal{A}_{M_I} \times \mathcal{A}_{M_I}$ are defeat relations for the object and meta-levels respectively.*
- *$\mathcal{S}_{MO_I} \subseteq \mathcal{A}_{M_I} \times \mathcal{A}_O$, is a support relation from meta-level to object-level arguments.*
- *$\mathcal{S}_{MA_I} \subseteq \mathcal{A}_{M_I} \times \mathcal{R}_O$, is a support relation from meta-level to object-level attacks.*

Bimodal argument graphs constrain arguments, requiring that for all $\phi \in \mathcal{A}_O$ and $(a, b) \in \mathcal{R}_O$ there exists a $\beta, \gamma \in \mathcal{A}_{M_I}$ such that $(\beta, \phi) \in \mathcal{S}_{MO_I}$ and $(\gamma, (a, b)) \in \mathcal{S}_{MA_I}$. If $(\beta, \phi) \in \mathcal{S}_{MO_I}$, then β is said to support ϕ.

Extensions within a bimodal argument graph (according to some semantics) are computed from the highest meta-level down to the object level. More specifically, the extension of the highest level is computed, and the subset of arguments at the next level down supported by arguments within the extension are used to form a sub-framework over which extensions are again computed. This process repeats itself until an extension at the object level can be computed.

3.2 The Object Level (l_0)

Our focus revolves around arguments obtained from a dialogue — a sequence of moves $D = [M_1, \ldots, M_x]$. We do not specify the protocol used to create this dialogue, but assume that each participant has a *commitment store* representing those arguments they are publicly committed to. Arguments can be *added* or *retracted* from each participant's commitment store. Furthermore, we assume that a participant is only committed to arguments that they have introduced. We denote the commitment store of participant Ag_i as CS_{Ag_i}, and call $\cup_{Ag_i} CS_{Ag_i}$ the *universal commitment store*, $\mathcal{UCS}$. The $\mathcal{UCS}$ corresponds to the set of arguments at the object level $\mathcal{A}_O$ in Def. 3. Both the individual and universal commitment stores are updated at each move of the dialogue.

Property	Definition
$\mathit{defeats}(a, b)$	argument a defeats argument b (i.e., $a, b \in \mathcal{A}$ and $(a, b) \in \mathcal{D}$)
$\mathit{unattacked}(a)$	argument a is unattacked (i.e., $a \in \mathcal{A}$ and $(b, a) \notin \mathcal{D}$)
$\mathit{preferred}(a, b)$	argument a is preferred to argument b (i.e., $a, b \in \mathcal{A}$, $(a, b) \vee (b, a) \in \mathcal{D}$ and a defeats b via meta-level arguments.
$\mathit{unattacked}(a, b)$	$\mathit{defeat}(a, b)$ is unattacked (i.e., $a, b \in \mathcal{A}$, $(a, b) \in \mathcal{D}$ and $(c, a) \notin \mathcal{D}$)
$\mathit{defended}(a, b)$	$\mathit{defeat}(a, b)$ is defended (i.e., $a, b, c, d \in \mathcal{A}$, $(a, b) \in \mathcal{D}$, $(c, a) \in \mathcal{D}$ and there is $(d, c) \in \mathcal{D}$)
$\mathit{conflict\_free}(CS_{Ag_i})$	the commitment store CS_{Ag_i} is conflict-free (i.e., there exist no $\phi_1, \phi_2 \in CS_{Ag_i}$ such that $(\phi_1, \phi_2) \in \mathcal{D}$)
$\mathit{retracted}(a, CS_{Ag_i})$	argument a is retracted from CS_{Ag_i} (i.e., $CS_{Ag_i} = CS_{Ag_i} \cup b$ and $(b, a) \in \mathcal{D}$)

Table 1: Predicates for Trust Properties

After introducing an argument at the object level, additional arguments are added to the meta-levels monotonically. Let $\varphi(\cdot)$ indicate that an element should be trusted. At the meta-levels, every argument $a \in \mathcal{A}_O$ is supported by an argument α asserting that a should be trusted ($\varphi(a)$), every defeat $(a, b) \in \mathcal{D}_O$ should also be trusted ($\varphi(a, b)$), and that utterances by an agent Ag_i should be trusted ($\varphi(Ag_i)$). Additional arguments are instantiated via trust-related argument schemes.

We map arguments and attacks in our hierarchical system to arguments and defeats in a bimodal argument graph [7] by stating that argument a defeats argument b iff a attacks b and there are some meta-arguments α, β such that α supports a and β supports b and α attacks β. Properties of the argument framework at the object level is encoded using a fragment of ASPIC+. We assume that $\mathcal{L}$ is a predicate-based language with a finite number of constant symbols, and which can therefore (formally) be mapped to a propositional language.

Agents build meta-arguments about object-level arguments, attacks, and sources of argument by applying a set of defeasible rules which we define as argument schemes (and critical questions). At the meta-level, we do not consider preferences between arguments, meaning that attacks and defeats are equivalent here.

3.3 The First Meta-level (l_1)

The first meta-level contains facts and associated rules from which arguments can be formed regarding the object level arguments. Table 1 summarises the predicates which can appear at the meta-level, and describes the condition under which these are added. As individual utterances are made within the dialogue, additional predicates and arguments are monotonically added to the meta-level. The arguments are obtained from a set of trust specific argument schemes. These schemes describe inference rules from which arguments can be created, as well as critical questions which allow attacks against the arguments to occur. We detail these schemes in the remainder of this section[3].

[3] Due to lack of space, we formalise only some of the schemes and critical questions.

Argument from Lack of Justification (Arg_{LJ}) If a dialogue participant cannot justify their arguments, then these arguments should not be trusted. More formally, if a is in Ag_i's commitment store, and b (in the universal commitment store) defeats a, then a is not (skeptically) justified. In turn, this means that the argument and dialogue participant should not be trusted. Formally, we have the following defeasible inferences.

$$r_{SLJ} : a \in CS_{Ag_i}, b \in UCS, \mathit{defeats}(b, a) \Rightarrow \neg\varphi(a)/\neg\varphi Ag_i$$

A defeater to b serves as a critical question to prevent the application of the scheme.

$$r_{CQLJ} : \exists c \in UCS, \mathit{defeats}(c, b) \Rightarrow \varphi(a)/\varphi(Ag_i)$$

Argument from Void Precedence (Arg_{VP}) This scheme is adapted from the void precedence property of ranking based semantics [1], and states that a non-attacked argument is accepted, and should therefore be considered trusted. We omit its formalisation due triviality and lack of space.

Argument from Defence Precedence (Arg_{DP}) This scheme is also adapted from ranking based semantics [1], and states that an argument defended against its attackers by more preferred argument(s) should be trusted.

$$r_{SDP} : a, b, c \in UCS, \mathit{defeats}(b, a), \mathit{defeats}(c, b) \Rightarrow \varphi(a)$$

At the same time, d defeating c would undercut this scheme, and serves as a critical question (not formalised due to space constraints).

Argument from Preference Precedence (Arg_{PP}) This scheme specifies how attacks between conflicting object-level arguments are resolved with preferences. In effect, an (otherwise defeated) argument which is preferred remains trusted as long as it is justified. Again, another defeater of the argument would render this scheme invalid.

$$r_{SPP} : a, b \in UCS, \mathit{defeats}(a, b), \mathit{preferred}(b, a) \Rightarrow \varphi(b)$$

Trust can be placed not only in arguments and speakers, but also in defeats. If we have $\{(a, b), (c, a)\} \subseteq \mathcal{D}$, then (c, a) attacks (a, b). An argument (d, c) would defend (a, b) in this case. A defeat is then trusted if it is unattacked, defended, or originates from a justified argument, and is untrusted otherwise. This intuition is also captured in extended argument frameworks with second (or higher) order attacks [5]. It should be noted that a defeat may be trusted when both arguments it refers to are untrusted. Argument schemes for reasoning about trust in defeats are defined as follows.

Argument from Justified Defeat (Arg_{JD}) A defeat is trusted if it originates from a justified argument.

$$r_{SJD} = a, b \in UCS, \mathit{defeats}(a, b) \Rightarrow \varphi(a, b)$$

As elsewhere, the presence of a defeater of a serves to undercut this scheme.

$$r_{CQJD} : c \in UCS, \mathit{defeats}(c, a) \Rightarrow \neg\varphi(a, b)$$

Argument from Unattacked Defeat (Arg_{UD}) A defeat is trusted if it is unattacked.

Argument from Defended Defeat (Arg_{DD}) A defeat is trusted if it is defended. This scheme is undercut if the defeat that the defender attacks is preferred to the defender.

3.4 The Second Meta-level (l_2)

In this level we consider properties that can be inferred to establish meta-arguments about trust in the sources of arguments at l_0. These meta-arguments indirectly attack or support arguments at level l_0 by attacking or supporting arguments at level l_1. For example, the assertion $\neg\varphi(Ag_i)$ (i.e., the source Ag_i of an argument a should not be trusted), attacks all meta-arguments at level l_1 which support arguments advanced by Ag_i at level l_0. Argument schemes here include the following.

Argument from Self Contradiction (Arg_{SC}) This scheme is adapted from Walton's argument from inconsistent commitment [12], and states that an agent committed to two arguments which attack each other should not be trusted.

$$r_{SSC} : a, b \in CS_{Ag_i}, \mathit{defeats}(a, b) \vee \mathit{defeats}(b, a) \Rightarrow \neg\varphi(Ag_i)$$

A closely related argument scheme is *Argument from Consistency* (Arg_{CN}) stating that if all an agent's commitments are conflict free, then the agent should be trusted.

Argument from Retraction (Arg_{RN}) Retracting a commitment results in a loss of trust. When performing such a retraction, some premises or warrants are also typically retracted [13]. This means that a retraction should cause trust to be lost not only for the retraction itself, but also for other arguments which are defended by the retracted argument (unless these latter arguments are defended by other unretracted arguments). This leads to the following scheme.

$$r_{SRN} : a, b \in CS_{Ag_i}, c \in UCS, \mathit{defeats}(c, a), \mathit{defeats}(b, c), \mathit{retracted}(b) \Rightarrow \neg\varphi(Ag_i)$$

We have described how meta-arguments about trust at different levels can attack each other and support lower level arguments. In our approach, each dialogue participant Ag_I has an associated BAG_I, whose object level is built from the dialogue and their commitment stores. Meta-levels components are constructed subjectively from a private knowledge base of preferences and properties observed at the object level. The maximal set of arguments appearing in the extensions of all participant's BAGs is the set of trusted arguments within the dialogue.

4 Example

Consider a long running dialogue between three agents (Ag_1, Ag_3, Ag_3) about the death penalty. At the object level, the following arguments are advanced.

- *Ag_1*: The death penalty is a legitimate form of punishment. (a)

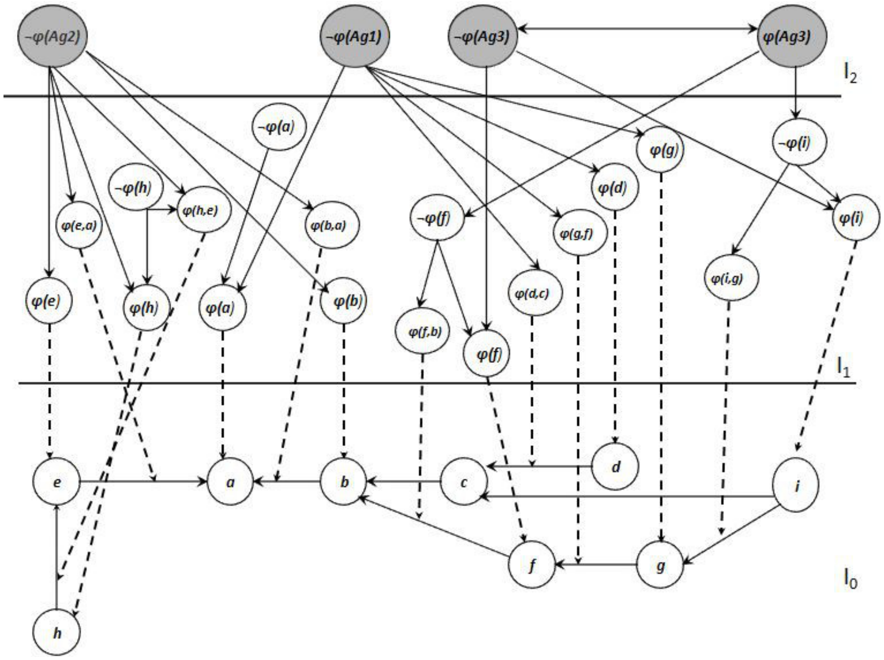

Fig. 1: Bimodal Graph for object and meta-level argumentation

- Ag_2: God does not want us to kill. (b)
- Ag_3: God does not exist. (c)
- Ag_1: Some people believe in God. (d)
- Ag_2: The state has no right to put its subjects to death. (e)
- Ag_3: The legal status of the death penalty should not depend on beliefs. (f)
- Ag_1: All religions should have a say over public law. (g)
- Ag_2: Majorities in some democratic countries favour death penalty .(h).
- Ag_3: Even if God exists, religion should stop at the door of the temple. (i)

Note that Ag_2 has potentially contradicted themselves in arguments e and h. Instantiating Arg_{SC}, we have an argument at the second meta-level for $\neg\varphi(Ag_2)$, which attacks $\varphi(e)$, $\varphi(h)$ and any other arguments advanced by Ag_2 in the dialogue. While argument h is undefeated, and supports argument a, yielding $\varphi(a)$ using Arg_{DP}, the fact that we had obtained $\neg\varphi(Ag_2)$ means that this support is attacked. Figure 1 provides the full bimodal argument graph obtained from this dialogue, where meta-arguments are represented by their conclusions.

5 Discussion and Conclusions

This paper presents an approach for reasoning about trust in dialogues that combines three of the most popular mechanisms used within computational modelling of argumentation: ASPIC+ [10], argument schemes [12] and meta-argumentation [7, 11].

Unlike the systems described in [2, 3] where preferences are given and fixed, our argument scheme based approach models how trust can be used as a rational basis for expressing preferences between arguments, determining successive attacks and for computing extensions. The systems in [8, 11] compute argument acceptability on the basis of the trustworthiness of their sources and the feedback that the final quality of arguments provide on the source evaluation. Unlike our approach, these approaches do not consider how trust in arguments and their sources change dynamically within a dialogue. Also the work presented in [9] has considered different argument schemes for reasoning about trust in an individual. However, these rely on extra-dialogical properties, while our focus is on how utterances affect trust during a dialogue.

We are pursuing several avenues of future work. First, we seek to link our system with graded and numerical semantics. Second, we recognise that the argument schemes we describe are not exhaustive, and believe that additional argument schemes for trust can be identified. Finally, we must demonstrate that the manner in which our system computes trust is consistent with human intuitions, and that it satisfies certain desirable properties. If divergences between these exists, then the framework could serve as a useful foundation for describing and studying paradoxes in human-based trust.

References

1. Amgoud, L., Ben-Naim, J.: Ranking-based semantics for argumentation frameworks. In: International Conference on Scalable Uncertainty Management. pp. 134–147. Springer (2013)
2. Amgoud, L., Vesic, S.: Rich preference-based argumentation frameworks. International Journal of Approximate Reasoning 55(2), 585–606 (2014)
3. Bench-Capon, T.J.: Persuasion in practical argument using value-based argumentation frameworks. Journal of Logic and Computation 13(3), 429–448 (2003)
4. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial intelligence 77(2), 321–357 (1995)
5. Modgil, S., Bench-Capon, T.: Integrating object and meta-level value based argumentation. Computational Models of Argument 172, 240–251 (2008)
6. Modgil, S., Prakken, H.: The ASPIC+ framework for structured argumentation: a tutorial. Argument & Computation 5(1), 31–62 (2014)
7. Müller, J., Hunter, A., Taylor, P.: Meta-level argumentation with argument schemes. In: International Conference on Scalable Uncertainty Management. pp. 92–105. Springer (2013)
8. Paglieri, F., et al.: Trusting the messenger because of the message. Computational and Mathematical Organization Theory 20(2), 176–194 (2014)
9. Parsons, S., et al.: Argument schemes for reasoning about trust. Argument & Computation 5(2-3), 160–190 (2014)
10. Prakken, H.: An abstract framework for argumentation with structured arguments. Argument & Computation 1(2), 93–124 (2010)
11. Villata, S., Boella, G., Gabbay, D.M., Van Der Torre, L.: A socio-cognitive model of trust using argumentation theory. Int. Journal of Approximate Reasoning 54(4), 541–559 (2013)
12. Walton, D., Reed, C., Macagno, F.: Argumentation Schemes. Cambridge (2008)
13. Walton, D., Krabbe, E.C.: Commitment in dialogue: Basic concepts of interpersonal reasoning. SUNY press (1995)
14. Wooldridge, M., McBurney, P., Parsons, S.: On the meta-logic of arguments. In: Proc. of the 4th Int. conference on Autonomous agents and multiagent systems. pp. 560–567 (2005)