Dynamic Trust in Dialogues

Gideon Ogunniye, Nir Oren, Timothy J. Norman
Department of Computing Science,
University of Aberdeen, AB24 3UE, Scotland, UK
g.ogunniye,n.oren,t.j.norman@abdn.ac.uk

Abstract

We present an argumentation framework for reasoning within deliberation dialogues. This framework includes a dynamic notion of trust — the degree of belief placed in the dialogue participants changes over the course of the dialogue, and in turn affects the strength of the arguments the participants advance, affecting the dialogue's conclusions.

Introduction

Within a dialogue, participants exchange advance arguments aimed at reaching some conclusion. Typically, these participants have partial information and individual preferences and goals, and the aim of the dialogue is for the parties to reach some outcome based on these individual contexts. Importantly, some dialogue participants may be *malicious* or *incompetent*, and the inputs from these parties should be discounted, based on the lack of trust ascribed to them.

While previous work (Paglieri et al. 2014) has considered how trust and reputation of participants should be updated following the justified conclusions of a dialogue, we observe that in long-lasting human discussions, trust can change during the dialogue itself. In turn, such changes in trust may require untrusted agents to present more evidence for their arguments to be believed, while the burden of proof reduces on highly trusted agents. Thus, there appears to be a feedback cycle which we would like to capture within more formal dialogue.

We therefore seek to address the following questions. 1) How can should trust affect the justified conclusions obtained from a dialogue? 2) How should trust change during the course of a dialogue based on the utterances made by the dialogue participants?

In the next section we provide a brief overview of abstract argumentation systems. Following this, we describe the components of our proposed system before discussing potential future work and concluding.

Background

Our system makes use of abstract argumentation, and we therefore begin by describing Dung's seminal approach (Dung 1995).

Definition 1 *An Argumentation Framework (AF for short) is defined as a pair* $\langle A, D\rangle$ *where A is a set of* arguments *and D is a binary defeat relation on A.*

Given an argumentation framework, one can identify different sets of justified conclusion by considering different *extensions*.

Definition 2 *Given an AF* $= \langle A, D\rangle$*, a set of arguments* $S \subseteq A$ *is said to be conflict-free iff* $\forall x, y \in S$*, there is no* $(x, y) \in D$*. Given an argument* $x \in S$*, S is said to defend x iff* $\forall y \in A$*: if* $(y, x) \in D$ *then there is a* $z \in S$ *such that* $(z, y) \in D$*.*

Then S is admissible iff it is conflict-free and defends all its elements. S is a complete extension iff there are no other arguments which it defends.

S is a preferred extension iff it is a maximal (with respect to set inclusion) complete extension. S is a grounded extension iff it is a minimal complete extension. S is a stable extension if it defeats all arguments not within S.

In this paper we will focus on the preferred semantics. These semantics admit multiple extensions; here, each such extension represents a potentially justified view (which conflicts with other views). If an argument is present in all extensions, then it is *sceptically* justified; while if it is present in at least one extension, it is *credulously* justified.

The System

We consider a system where dialogue participant i is modelled through a *commitment store* $CS_i \subseteq A$, containing a set of arguments. At any point in time, a participant may *add* or *retract* arguments from their commitment store. An argument may be added to a commitment store if it is not already present within it (and was not previously present), and may be retracted only if it was already present. We also consider the *universal commitment store* $UCS = \bigcup_i CS_i$. The dialogue then consists of a sequence of *add* and *retract* moves, where each move references both an argument and a dialogue participant (e.g., $add(\alpha, a)$ denotes that a adds an argument α to their commitment store.

Each dialogue participant also has an associated trust rating, encoded through a preference ordering over all dialogue participants (this preference ordering is represented by the relationship $\succeq$). Given a universal commitment store, a set

of *attacks* between arguments[1], and a preference ordering over dialogue participants, we can instantiate an abstract argumentation framework by transforming attacks into defeats as done in (Modgil and Prakken 2012): argument a defeats an argument b iff a attacks b and there are some dialogue participants α, β such that $a \in CS_\alpha, b \in CS_\beta$ and $\alpha \succeq \beta$.

Our approach is based on the following observations.

- A dialogue participant whose arguments are self-contradicting should be less trusted than a consistent participant.
- A dialogue participant who is unable to justify their arguments should be less trusted than one who can.
- A dialogue participant who regularly retracts arguments should be less trusted than one who does not.

We seek to formalise each of these observations within our framework.

Self Contradicting Arguments

A dialogue participant i is self contradicting iff there are two arguments $a, b \in CS_i$ such that a attacks b or vice-versa.

We write SC_i to denote the number of self contradicting arguments dialogue participant i has.

Lack of Justification

There are several ways to formalise a lack of justification. For example, one could consider partial arguments, though these should be distinct from enthymemes. Given the abstract nature of our system, we consider unjustified arguments as those that are defeated, but not reinstated (i.e., those that do not appear within the extension according to the semantics under which the dialogue operates).

Formally, an argument a lacks justification for a dialogue participant i iff $a \in CS_i$ and $a \notin \mathscr{E}(UCS, D)$. Here, $\mathscr{E}$ represents the extension(s) obtained on the argumentation framework $\langle UCS, D \rangle$.

We note again that additional definitions of the lack of justification are possible. For example, one could require that the dialogue participant in question be the one to advance the reinstating argument.

We write LJ_i to denote the number of arguments associated with dialogue participant i that lack justification.

Argument Retraction

The number arguments retracted by an dialogue participant i is denoted AR_i.

Computing Trust

At any point in the dialogue, we may compute SC_i, LJ_i and AR_i for every agent. Then we may compute a *trust value* for each dialogue participant according to a function $trust : \mathbb{Z} \times \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$. This trust value provides us with a total order — those with trust value 0 are most trusted, while higher (numerical) trust values are less trusted. In turn, this total order is used to compute defeats at that point in the dialogue.

[1] Like defeats, an attack is a binary relation over arguments.

Discussion

We have not proposed a specific *trust* function. Different dialogue participants (or external observers) may wish to weight the different factors affecting trust differently, with some for example finding self contradiction most important, while others may find lack of justification, or a combination of factors more relevant[2].

Trust may change at any point in the dialogue, in turn affecting the defeat relationship. Such changes can cause arguments to disappear from an extension, affecting not only the conclusions of the dialogue, but also the trust placed in dialogue participants. This change in trust can then further affect arguments, leading to a vicious, or virtuous, cycle. This leads to an important questions which we are currently investigating, namely under what conditions our system is stable, and whether these conditions agree with common-sense intuitions. A different way of instantiating the system, with no cycles, requires changes in trust to affect the defeat relation in the following, rather than past and present dialogue states. We therefore need to determine which of these two approaches is more appropriate.

Another avenue of future work involves identifying realistic trust functions. Furthermore, the three observations used to modify trust are by no means exhaustive, and we wish to investigate additional factors that not only decrease trust, but may also increase it. Finally, extending this work to instantiated argumentation systems, and specific dialogues may yield additional factors which affect trust, which should also be investigated.

Conclusions

In this paper we described a system in which the arguments advanced, or retracted, by a dialogue participant affects the trust placed in them. In turn, this trust affects trust in the participant's arguments, which may lead to different conclusions being drawn.

We described three factors which modify trust, and how extensions can be computed within such a system. This work is preliminary, and we also identified a research path we are currently pursuing to create a complete system and understand its properties.

References

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

Modgil, S., and Prakken, H. 2012. A general account of argumentation and preferences. *Artificial Intelligence Journal* 361–397.

Paglieri, F.; Castelfranchi, C.; da Costa Pereira, C.; Falcone, R.; Tettamanzi, A.; and Villata, S. 2014. Trusting the messenger because of the message: feedback dynamics from information quality to source evaluation. *Computational & Mathematical Organization Theory* 20(2):176–194.

[2] Perhaps the simplest trust function involves simply summing up the three features.