Generating Discourse-Based Explanations

Andrew Potter

Humans and artificial agents need to be able to explain themselves to one another. They need to be able to present their perspectives and assess the views of others. This paper describes an approach to explanation aware reasoning, using underlying structures of natural discourse and argumentation theory. By positioning argumentation, explanation, and defeasibility concepts as first-class ontological entities, we can create an explanation-aware solution for multi-agent environments. Arguments are linked using ontologically specified interactions, such as substantiation, rebuttal, and accrual. It is envisioned that communities of human and artificial agents will engage in collaborative explanatory argumentation using argumentative structures and interactions for discovering knowledge and managing and navigating conflict and agreement.

1 Introduction

Humans and artificial agents need to be able to explain themselves to one another. However, providing a facility for explanation has not come easily, particularly with respect to explanation services for multi-agent systems. What is needed is a general theory of explanatory reasoning for human-MAS collaboration. As a step in this direction, we describe a discourse-based argumentative reasoning theory intended for support of intelligent human-computer collaboration. By positioning argumentation, explanation, and defeasibility concepts as first-class ontological entities, we can create an explanation-aware solution for multi-agent environments. We are developing an approach based on underlying structures of natural discourse and argumentation theory. Arguments, when satisfied, are instantiated into rhetorical networks that represent an agent's model of the situation. Instantiated arguments relate to one another using ontologically specified rhetorical interactions. The rhetorical network provides a coherent model which may be rendered graphically. These networks are similar to the inference networks used by Pollock [1] and the explanation paths proposed by Wærn and Ramberg [2]. Communities of human and artificial agents will engage in collaborative explanatory argumentation using these interactions as mechanisms for handling conflict and agreement.

This paper reports on an ongoing project introduced in an earlier study [3] on explanation-aware computing. We build on those concepts by applying them to the problem of explanation to a real-world problem, and we extend our earlier work by providing a fuller development of the concept of a rhetorical network.

2 Theoretical Background

The approach described here uses the Toulmin model and Rhetorical Structure Theory (RST) as the basis for its ontology. The Toulmin model is a theory of argumentation [4]. Toulmin defined an argument as consisting of six elements: a claim, a ground, a warrant, a backing, a qualifier, and a rebuttal. The claim is what the argument seeks to demonstrate. The ground is the data that support the claim. The warrant establishes the linkage between ground and claim. The backing is a policy, law, argument, or fact that substantiates the warrant. The qualifier is an indication of the strength of the argument. The rebuttal is any counter-argument that might refute the argument.

Rhetorical Structure Theory is a theory of text coherence [5]. RST defines the coherence of a text in terms of the way its parts, or text-spans, relate to one another. It postulates a small number of schemas for defining the possible structural relationships among spans and defines a set of rhetorical relations for use in applying a schema to a set of text spans. Generally, a schema will specify a nucleus and one or more satellites, where the satellite text-spans are in a supportive role to the nucleus. An RST analysis of a coherent document defines a hierarchical structure representing the rhetorical interrelationships of the text spans comprising the document.

3 Argumentative Ontology

The argumentative ontology defined in our earlier report [3] specifies a conceptualization for representing and managing argumentative and explanatory knowledge. As a theory of how the constituents of an argument interrelate, it integrates the Toulmin model with RST. Generally, Toulmin's grounds and claims correspond with RST satellites and nuclei. RST supplies relations for describing the nature of the relationship between ground and claim. To realize this, it is necessary to formulate these concepts ontologically. Arguments thus specified are not merely a means of affirming claims on the basis of grounds; they are objects of knowledge and may be handled accordingly.

In the ontology, an Argument defines a warrant and a set of Interactions. The Warrant defines the Nucleus (Claim) and the Satellite (Ground). The Nucleus is a Statement, which is an ontologically normalized expression, usually of a domain specific nature. The Satellite specifies the satellite Statement and its Relation to the Nucleus. The Relation identifies the RST relation and characterizes its modality as either synthetic or inferential. The modality of an argument constrains the conditions under which it may be instantiated. The Argu-

Auszug aus: Künstliche Intelligenz, Heft 2/2008, ISSN 0933-1875, BöttcherIT Verlag, Bremen, **www.kuenstliche-intelligenz.de/order**

ment also identifies a qualifier, a qualification ratio, and a set of argument interactions. The qualifier is an a priori value indicating the level of certainty of an argument, and may be either conclusive or supportive. The qualification ratio is dynamic and is defined by the interactions in play between an instantiation of the argument and other argument instantiations. Interactions are the relations among arguments in the rhetorical network.

4 Rhetorical Networks

To understand a situation is not simply to comprehend a collection of discrete facts and inferences, but to realize how these facts and inferences interact with one another to produce an integrated model. This model is an explanation. Rhetorical networks provide the means to represent integrated models arising from multiple viewpoints. A rhetorical network is a coherent representation of a situation. The network consists of argument instantiations and interactions among these instantiations. An instantiation is an application of an argument. In the following cyber defense example, it is argued that the event of a compromised is sufficient motivation to quarantine the host from the network:

```
argument(
 warrant(
   satellite(compromised( host( T)),
     relation( motivation, inferential )),
   nucleus(recommendation(quarantine(host(T))))),
 qualifier( supportive ),
 qr( QR ),
 interactions( I ))
```

The argument is satisfied when the satellite is satisfied. Instantiation of the argument binds its variables and asserts the instance into the rhetorical network. Note that what is asserted is not simply the claim, but the entire argument. This is important in producing explanations, as will be discussed in the next section.

There are two modalities of instantiation. Inferential instantiation, used in the previous example, applies to arguments whose claims may be inferred from their grounds. Instantiation of the ground is sufficient to instantiate the argument. Synthetic instantiation is applicable to arguments where both ground and claim must be satisfied for the argument to be instantiated. Here is an example of a synthetic argument:

```
argument(
 warrant(
   satellite(procedure(quarantine, host(T)),
     relation( enablement, synthetic )),
   nucleus(recommendation(quarantine( host(T))))),
 qualifier( conclusive ),
 qr( QR ),
 interactions( I ))
```

This argument establishes a linkage between a recommendation and the procedure for carrying out the recommendf. Thus, the role of synthetic relations is to provide additional information.

An argument, once instantiated, may interact with other arguments. They may conflict with one another by disputing each other's claims, their grounds, or their warrants. They may converge upon a single claim, or, from a single ground, multiple claims may issue forth. In any interaction, each argument has a locus of interaction which identifies the elements engaged in the interaction. Some interactions engage claims, some grounds, some combinations grounds and claims, and some combinations of claims and warrants.

Interactions are typically, but not always, asymmetric. That is, one argument attempts to influence another, but the influence is not reciprocated . For example, one argument that substantiates another is not itself substantiated by the latter (unless the reasoning is circular). Thus, in an argument interaction, one argument may be designated as the catalyst and the other as the reactant. If the influence exerted by the catalyst is supportive, the polarity is positive. If the influence is resistant, the polarity is negative. In some interactions, the influence may be benign, or neutral. When one argument substantiates another, the catalyst exerts a positive polarity on the reactant. When an argument rebuts another, the catalyst exerts a negative polarity on the reactant. Polarity is used in detecting when two arguments are in conflict with one another. These features may be used as hints when reasoning about interactions.

4.1 Substantiation and Rebuttal

When the claim of one argument unifies with the ground of another, the *substantiation* interaction is specified. Substantiating arguments may be syllogistically chained to one another through a claim-ground linkage to form a logical sorites. In the following definition, the claim, C1, of the first argument is also the ground of the second:

```
substantiation(arg(G1,C1,W1) & arg(C1,C2,W2))
```

In *rebuttal*, the loci of interaction are in the claims made by the interacting arguments. One claim disputes the other. The incompatibility may be either logical or ontological, the polarity is negative, with the catalyst is in the rebutting argument, and the reactant is the argument subjected to rebuttal. Rebuttal is structurally symmetrical, such that if argument *p* rebuts *q*, then *q* rebuts *p*:

```
rebuttal(arg(G1,C1,W1) & arg(G2,C2,W2))
  & claim(incompatible(C1,C2))
```

4.2 Backing, Undercut, and Dissociation

In Toulmin theory, *backing* is the policy, law, argument, or fact that supports the warrant. More generally, we may say that backing is any argument that substantiates a warrant. It is an argument with positive polarity whose locus of interaction resides in the claim of the catalyst and the warrant of the reactant. The claim of one instantiated argument substantiates the warrant of another:

```
backing(arg(G1,claim(W2),W1) & arg(G2,C2,W2))
```

Some researchers [1,6] have used the term *undercut* to refer to a claim that challenges a warrant. As used here undercut refers to a less disruptive form of challenge, where

the claim of one instantiated argument simply disputes the ground of another:

```
undercut(arg(G1,claim(C1),W1)
  & arg(ground(G2),C2,W2)
  & incompatible(C1,G2))
```

Dissociation is used to denote the more disruptive form of interaction, when the claim of one argument disputes the warrant of another. Dissociation challenges the standing of the argument itself. An argument, once dissociated, is no longer an argument. The ground is no longer a ground and the claim is no longer a claim; rather they are henceforth dissociated units:

```
dissociation(
  arg(G1,claim(C1),W1)
  & arg(G2, C2, W2)
  & incompatible(C1,W2))
```

4.3 Accrual, Concomitance, and Confusion

In *accrual*, multiple arguments lead to the same claim, or multiple instantiations of the same argument lead to an identical claim:

```
accrual(arg(G1,C1,W1) & arg(G2,C1,W2))
```

Intuitively it might seem that accruing arguments would collectively strengthen the claim. However, there seems to be no means for abstractly evaluating accrual, as has been pointed out by several researchers [6-7]. The evaluation must be ontologically specific.

Concomitance occurs when two arguments use the same ground to establish distinct claims. Concomitance is non-catalytic, and the polarity is neutral.

```
concomitance(arg(G1,C1,W1)
  & arg(G1,C2,W2))
```

Finally, *confusion* occurs when incompatible grounds are instantiated. If either of the grounds is also the claim of some other argument, the condition may more appropriately be handled as an undercut; otherwise, confusion prevails.

```
confusion(arg(G1,C1,W1)
  & arg(G2,C2,W2)
  & incompatible(G1,G2))
```

5 Explanations

Argumentation and explanation are closely allied discourse modalities. The principal distinction is that in argumentation, there is a presumption that the claims presented may not comprise the sole possible interpretation of a situation [8-9]; multiple points of view are possible. This is an important accommodation in multi-agent environments. The explanatory power of arises not only from the use of the rhetorical model, but through interaction among inferential and synthetic interaction structures. By this means, explanatory information

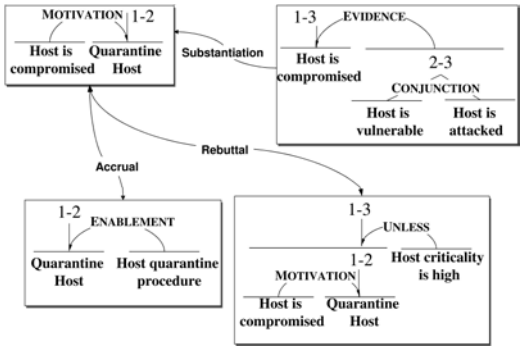

Figure 1: Rhetorical Network

is contained in the interaction structure. For example, a substantiation structure may be inferred when its catalyst and reactant claims are satisfied. The substantiation then functions as a structured explanation for the reactant argument. In Figure 1, an argument offering evidence that a computer host has been compromised offers substantiation for the argument that the host should be quarantined.

Further, any claim may be augmented with synthetic support. In Figure 1, the argument offering a quarantine procedure as Enablement for quarantining a host is synthetic. Because this argument and the argument for quarantining the host share the same nucleus, they relate in an accrual interaction.

Because the explanations are argumentative, the rhetorical network may also contain conflicting information. Thus, in Figure 1, a rebuttal has been added to the network, indicating that perhaps the quarantine should not be carried out. In this way, the rhetorical network provides a representation of the system's model of the situation.

Further, it is possible to use rhetorical networks to produce a synthesis of the interacting arguments, as shown in Figure 1. Here the network has been reduced to a single argument consisting of multiple warrants and claims. This is accomplished through unification of the claims. The synthesis produces a structured "paragraph," or complex argument, describing conditions, policies, and their interactions. Like conventional RST analyses, synthesized rhetorical networks are subject to the constraints of completeness and connectedness. Unlike conventional RST analyses, however, rhetorical networks are not subject to the constraints of uniqueness and adjacency. This permits them to be used to model richly interconnected interactions. Rhetorical network structures are defined in terms of two schemas, the satellite-nucleus and nucleus-satellite schemas. Both schemas consist of a satellite, nucleus, and a relation. The schemas imply an ordering. In the satellite-nucleus schema, the satellite precedes the nucleus, and in the nucleus-satellite schema, the nucleus precedes the satellite. Graphically, applications of the satellite- nucleus are represented with the satellite above the nucleus and the arrow pointing downward. Applications of the nucleus-satellite schema are represented as a satellite below the nucleus with an arrow pointing upward from the satellite to the nucleus.

Some relations are associated with a specific schema type [5]. The associations are based on the implied temporal considerations of the relation. For example, the Evaluation uses

Auszug aus: Künstliche Intelligenz, Heft 2/2008, ISSN 0933-1875, BöttcherIT Verlag, Bremen, **www.kuenstliche-intelligenz.de/order**

the nucleus-satellite schema because the satellite must follow the nucleus if it is to evaluate it. In Figure 1, the Motivation relation uses the Satellite-Nucleus schema, and the Evidence, Enablement and Unless relations use the nucleus-satellite schema. In addition to providing explanations, the paragraphs produced using network synthesis may be useful in identifying multiple situations as similar or analogous. It is possible that partial isomorphism between complex arguments could be useful in knowledge discovery.

6 Summarization heuristics

The ability to summarize rhetorical networks could be used in simplifying information that is presented to the user or to other agents. In a rhetorical structure, nuclear segments are more central to the argument than their satellites. That this attribute of RST lends itself to automatic summarization is a point made by the theory's originators [5]. However, mechanically pruning off outer vertices is unlikely to lead to satisfactory results-crucial information could be lopped off in the process. Summarization heuristics can be used for a more refined process. For example, satellites of synthetic arguments might be pruned more liberally than satellites of inferential arguments, and, under some circumstances, arguments with a positive polarity might be pruned before those with a negative polarity. The use of qualification ratios (discussed below) will also play a role in summarization by providing a means for eliminating weaker lines of reasoning.

7 Representing uncertainty

Our preliminary approach to determining argument strength is that the methodology be simple and intuitive. The strength of a claim can be measured as a qualification ratio: the higher the ratio, the more established the claim [10]. The qualification ratio is a measure of the certitude of the total argumentative structure. This ratio is derived from the qualifiers associated with the interacting arguments. Given two competing arguments, the one with the higher qualification ratio is more certain.

8 Conclusion

The concepts presented here provide some essential elements for discourse based explanation generation. These concepts may be reduced to a few key ideas. First, reasoning results in instantiation of argumentative structures rather than discrete propositions. The instantiations arise as a result of input from a variety of sources, including both human and artificial agents. Given a set of instantiations, we can locate the negative and positive interactions among them, and use these interactions to define rhetorical networks of argument structures. By unifying the loci of each interaction, we may synthesize the networks into coherent rhetorical expressions, or paragraphs. These paragraphs are argumentative explanations of a situation, as known to an agent. The organization of paragraphs is based on underlying structures of natural discourse and argumentation theory. Whereas the organization of paragraphs seems to be domain neutral, the statements contained within these structures are ontologically normalized expressions. It is envisioned that communities of human and artificial agents will engage in collaborative explanatory argumentation using argumentative structures and interactions for discovering knowledge, detecting, managing and navigating conflict and agreement. Future work will include additional research in argument interaction, particularly with respect to accrual, analogical reasoning, summarization, and development of algorithms for management of qualification ratios.

References

[1] J. L. Pollock, Cognitive carpentry: A blueprint for how to build a person. Cambridge, MA: MIT Press, 1995.

[2] Y. Wærn and R. Ramberg, "Distributed knowledge by explanation networks," in HICSS '04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 5 Washington, DC: IEEE Computer Society, 2004.

[3] A. Potter, "A discourse approach to explanation aware knowledge representation," in AAAI 2007 Workshop on Explanation-aware Computing, T. Roth-Berghofer, S. Schulz, D. B. Leake, and D. Bahls, Eds. Vancouver, British Columbia, in press.

[4] S. E. Toulmin, The Uses of Argument. Cambridge, UK: Cambridge University Press, 1958.

[5] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Towards a functional theory of text organization," Text, vol. 8, pp. 243-281, 1988.

[6] H. Prakken, "A Study of accrual of arguments, with applications to evidential reasoning," in The Tenth International Conference on Artificial Intelligence and Law, Proceedings of the Conference, June 6-11, 2005 New York: ACM, 2005, pp. 85-94.

[7] R. J. Yanal, "Linked and convergent reasons - Again," in Informal logic at 25: Proceedings of the Windsor Conference, J. A. Blair, R. H. Johnson, H. V. Hansen, and C. W. Tindale, Eds. Windsor, ON, 2003.

[8] D. Walton, "Examination dialogue: An argumentation framework for critically questioning an expert opinion," Journal of Pragmatics, vol. 38, pp. 745-777, 2006.

[9] B. Moulin, H. Irandoust, M. Bélanger, and G. Desbordes, "Explanation and argumentation capabilities: Towards the creation of more persuasive agents," Artificial Intelligence Review, vol. 17, pp. 169-222, 2002.

[10] J. Fox and S. Das, Artificial intelligence in hazardous applications. Menlo Park, CA: AAAI Press, 2000.

Contact
Dr. Andrew Potter
Sentar Incorporated
4900 University Square Suite 8
Huntsville, Alabama 35816
Tel.: 00 + 1 256-430-0860
Fax: 00 + 1 256-430-0840
Email: apotter@sentar.com

Andrew Potter is a research scientist with Sentar, Incorporated in Huntsville, Alabama, USA. His current and recent research interests include ontological knowledge representation, explanation aware computing, situation awareness systems for cyber defense, and asynchronous learning environments.