II Jornadas de Investigación en Humanidades

30, 31 de agosto y 1 de septiembre 2007

Universidad Nacional del Sur
Departamento de Humanidades
Bahía Blanca, Argentina

Auspiciantes:

Fundación Ezequiel Martínez Estrada

Cátedra Libre de Derechos Humanos del Departamento de Humanidades de la Universidad Nacional del Sur

Herramientas formales en el estudio de la argumentación rebatible

Gustavo Adrián Bodanza
Universidad Nacional del Sur/CONICET
ccbodanz@criba.edu.ar

1. Argumentación rebatible

Argumentar rebatiblemente es construir razonamientos mediante inferencias no concluyentes, aceptando o rechazando las conclusiones en virtud de las razones que las soportan. La argumentación rebatible se manifiesta cotidianamente en todos los aspectos de la vida social. Un ejemplo concreto está dado por los procesos judiciales, donde debe dirimirse una sentencia sopesando argumentos a favor y en contra de un cargo. Desde mediados de los añps '80 el estudio de la argumentación ha sido impulsado como línea de investigación en la representación del razonamiento de sentido común, y todavía sigue vigente la propuesta de distintos modelos formalizados.

Respondiendo a la pregunta sobre qué se considera un argumento rebatible y cómo se lo representa, es usual ver en los diversos sistemas dos aspectos diferenciados, uno *lógico* y otro *dialéctico*. El aspecto *lógico* es aquél bajo el cual se define la construcción de un argumento rebatible, su estructura y sus componentes, así como las reglas que permiten tal construcción. En los procesos jurídicos, por ejemplo, este aspecto sería el que determina cuáles son los argumentos atingentes para el caso en cuestión teniendo en cuenta el código de aplicación correspondiente. El aspecto *dialéctico* comprende a su vez los criterios que hacen a unos argumentos *preferibles* a otros, los que hacen que unos argumentos *rebatan* o *derroten* a otros, y los que determinan cuáles argumentos son los *ganadores* o *justificados* a lo largo del proceso dialéctico en el cual interactúan. Siguiendo el ejemplo anterior, bajo este aspecto se determina cuáles son las normas más apropiadas entre las aplicables al caso y, entre ellas, cuáles se imponen (por ejemplo, normas de mayor jerarquía, como las constitucionales, se imponen sobre aquellas de códigos inferiores en caso de conflictos). Desde el punto de vista de la Inteligencia Artificial, un tercer aspecto importante de la argumentación rebatible es el de los *procesos*, es decir, el de encontrar los algoritmos necesarios para implementar la argumentación rebatible. En este trabajo obviaremos este aspecto ya que requiere ciertos detalles técnicos especiales.

El problema que nos ocupa es cómo formalizar los aspectos lógico y dialéctico de la argumentación rebatible. Brevemente, el aspecto lógico es usualmente formalizado a través de un lenguaje lógico de primer orden, ampliado mediante la introducción de reglas de inferencia que expresan razones rebatibles o *prima facie* (Reiter, Lin & Shoham, Loui, Pollock, etc.). El aspecto dialéctico, en cambio, no puede ser —al menos hasta el momento no ha sido— reducido a un sistema puramente lógico, pero sí puede ser descripto utilizando herramientas de lenguaje matemático (Pollock, Dung). Presentaremos, entonces, algunas de las herramienta formales que permiten modelar un marco argumentativo, con el cual pueden representarse los argumentos usados en una discusión y las interacciones entre estos, para modelizar luego la selección final de los mejores argumentos. Sólo supondremos del lector un mínimo conocimiento de la lógica de primer orden.

2. Lenguaje de argumentos

En primer lugar, un vamos a considerar informalmente a un "argumento" como una entidad compuesta por proposiciones de las cuales una de ellas, la "conclusión", está en cierta conexión lógica con las otras, las "premisas". Para la formalización de este concepto comenzaremos construyendo un lenguaje los lineamientos principales de Pollock (1990). Sea $\mathsf{L}_A = \langle \mathsf{L}, \Delta \rangle$ un lenguaje que llamaremos *lenguaje argumentativo*, formado por un lenguaje de primer orden L, y un conjunto Δ de *reglas prima facie.* Las reglas prima facie serán entidades metalingüísticas respecto de L, y tendrán la forma '$\alpha \succ\!\!— \beta$', siendo informalmente interpretadas como 'creer α es una buena razón para creer β'. Los téminos de las reglas en Δ serán variables y valdrá la instanciación mediante constantes. Las fórmulas del lenguaje L_A son *argumentos*, estructuras de la forma '$\langle Def, Sup, Con \rangle$' donde $Def \subseteq \Delta$ es el *soporte rebatible* del argumento, $Sup \subseteq \mathsf{L}$ son los *supuestos* del argumento, y $Con \in \mathsf{L}$ es la *conclusión* del argumento. Un argumento es *suposicional* si $Sup \neq \varnothing$, de otro modo es *fundado*; y es *rebatible* si $Def \neq \varnothing$, de otro modo es *conclusivo*.

Finalmente, '$\Rightarrow$' es una relación de *consecuencia argumentativa* que permite construir un argumento a partir de otros. Esta relación está definida por el siguiente conjunto de reglas:

Suposición. Para cualquier conjunto $Sup \subset \mathsf{L}$ y cualquier $Con \in Sup$, $\varnothing \Rightarrow \langle \{\}, Sup, Con \rangle$.

Deducción. Si $\{C_1, ..., C_n\} \vdash Con$ entonces $\{\langle Def_1, Sup_1, C_1 \rangle, ..., \langle Def_n, Sup_n, C_n \rangle\} \Rightarrow \langle Def_1 \cup ... \cup Def_n, Sup_1 \cup ... \cup Sup_n, Con \rangle$.

Modus ponens rebatible. Si $C \succ\!\!— Con \in \Delta$ entonces $\{\langle Def, Sup, C \rangle\} \Rightarrow \langle Def \cup \{C \succ\!\!— Con\}, Sup, Con \rangle$ (donde en '*C*' ocurren las mismas constantes que en '*Con*').

Condicionalización. $\{\langle Def, Sup \cup \{C\}, Con \rangle\} \Rightarrow \langle Def, Sup, (C \rightarrow Con) \rangle$ (donde '$\rightarrow$' es el condicional material clásico).

De esta manera podemos construir todos los argumentos bien formados en un contexto argumentativo determinado. Por *contexto argumentativo* vamos a entender un par $CA = \langle BA, \Rightarrow \rangle$, donde *BA* es la *base argumentativa*, un conjunto arbitrario de argumentos, con el que podemos representar los argumentos básicos de un agente (o de varios agentes, si se tratara de representar una discusión un debate). A partir de los argumentos de la base argumentativa y aplicando las reglas que rigen la relación $\Rightarrow$ podrán encontrarse todos los argumentos correctos en el contexto.

3. Sistema

Ejemplo 1.
Supongamos un contexto formado por una base argumentativa con los siguientes argumentos:
1. $\langle \{\}, \{murciélago(\text{Bruno})\}, murciélago(\text{Bruno}) \rangle$
2. $\langle \{\}, \{murciélago(x) \rightarrow mamífero(x)\}, murciélago(x) \rightarrow mamífero(x) \rangle$

y donde las reglas prima facie son:

Δ= { r1: *mamífero*(x) $\succ\!\!-$ $\neg$*vuela*(x),
r2: *murciélago*(x) $\succ\!\!-$ *vuela*(x)}.

Entonces podemos construir los siguientes argumentos correctos:

3. $\langle$\{*mamífero*(Bruno) $\succ\!\!-$ $\neg$*vuela*(Bruno)\}, \{*murciélago*(Bruno)\}, $\neg$*vuela*(Bruno)$\rangle$ (de 1 y 2 aplicando r1)
4. $\langle$\{*murciélago*(Bruno) $\succ\!\!-$ *vuela*(Bruno)\}, \{*murciélago*(Bruno)\}, *vuela*(Bruno)$\rangle$ (de 1 aplicando r2)

Podemos ver cómo en un contexto de creencias es posible construir argumentos rebatibles a favor y en contra de una misma tesis. Lo que aún no podemos decidir es cuál de los argumentos opuestos es el "justificado" o elegido. Esto, que es parte del aspecto dialéctico de la argumentación rebatible, es lo que trataremos a continuación.

4. Dialéctica

El primer punto a considerar desde el punto de vista dialéctico es el de establecer un orden de preferencias sobre los argumentos. Los criterios pueden variar dependiendo materialmente del cuerpo de creencias sistematizado. Por ejemplo, si un sistema representa un código de justicia en el contexto de una acción legal particular, un buen criterio puede ser el que hace preferibles aquellos argumentos que aplican las normas más específicas para el caso en cuestión. El criterio también sería útil para dirimir el problema del ejemplo 1: el argumento *arg2* es preferible al argumento *arg1* puesto que el primero aplica una regla que se ajusta al hecho de que Bruno es específicamente un murciélago, mientras el segundo se basa en el hecho más general de que Bruno es un mamífero. Otros criterios de preferencia pueden basarse en la plausibilidad de la información particular (premisas preferidas), en la plausibilidad de las reglas según la fuerza con que éstas conectan el antecedente con el consecuente, según una preferencia entre las reglas mismas (e.g., en el campo jurídico, los principios *lex specialis derogat legi generali*, *lex superior derogat legi inferiori*, o *lex posterior derogat legi priori*.), etc. Cuando existe un único criterio de preferencia, éste simplemente puede utilizarse para establecer un orden parcial sobre los argumentos. Pero cuando los criterios son múltiples es preciso definir una jerarquía (por ejemplo, a través de un orden lexicográfico).

Desde el punto de vista formal, las propiedades de las preferencias han sido bien estudiadas, sobre todo en el campo de la economía. Las mismas usualmente se consideran reflexivas y transitivas, dando lugar a pre-ordenamientos de los elementos comparados. Pero cuanto más fuertes son los presupuestos de racionalidad de los agentes, otras propiedades se agregan. Por ejemplo, puede considerarse que el agente es capáz de ordenar completamente las opciones de acuerdo a sus preferencias en el sentido de que para cada par de elementos siempre hay un tercer elemento comparable o bien con el primero o bien con el segundo. Simari & Loui (1992) establecen una relación preferencia entre argumentos según la *especificidad* de los mismos (basada a su vez en la Poole (1987)) que desde el punto de vista formal cumple con las propiedades reflexiva, antisimétrica y transitiva. Según este cirterio, el argumento 3 de nuestro ejemplo 1 es al menos tan específico como el argumento 4, pero no a la inversa.

El criterio de especificidad puede verse claramente mediante una representación de los argumentos como grafos dirigidos. Los nodos del grafo representan proposiciones, y las flechas muestran cómo estas proposiciones están conectadas por reglas. Las flechas sólidas

representan condicionales materiales en sentido clásico ("si Bruno es un murciélago, entonces *necesariamente* Bruno es un mamífero"), mientras las flechas punteadas indican conexiones tentativas mediante reglas prima facie (" si Bruno es mamífero, entonces *tentativamente* Bruno vuela"). Una cruz sobre una flecha indica la conexión con la negación de la proposición señalada. En la fig. 1 pueden verse representados a la vez los argumentos 3 y 4 del ejemplo 1.

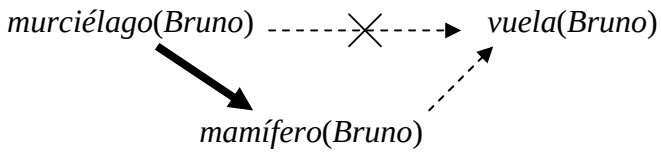

Figura 1

Con la ayuda de este grafo, el criterio de especificidad puede enunciarse así: el argumento *arg1* es *al menos tan específico como* el argumento *arg2* si y sólo si:
1. Todo nodo que —siguiendo el sentido de las flechas— conecta con la conclusión de *arg1*, conecta también con la conclusión del argumento *arg2*.
2. Existe un nodo que conecta con la conclusión de *arg2* pero no conecta con la conclusión de *arg1*.
Así podemos ver que en nuestro ejemplo el argumento 3 es al menos tan específico como el argumento 4 ya que, todo nodo que conecta con '¬*vuela*(*Bruno*)' —la conclusión de 3—, a saber, '*murciélago*(*Bruno*)', conecta también con '*vuela*(*Bruno*)' —la conclusión de 4. Y existe un nodo que conecta con '*vuela*(*Bruno*)' —la conclusión de 4—, a saber, '*mamífero*(*Bruno*)', pero no conecta con '¬*vuela*(*Bruno*)' —la conclusión de 3. Es claro que la inversa no se da.

Como puede verse hasta aquí, tenemos una representación de los argumentos usando las herramientas usuales de la lógica, pero la representación de la interacción argumentativa se ha visto facilitada por la utilización de nociones extra-lógicas —p. ej. relaciones binarias y sus propiedades y modelos representacionales como grafos.

1. Derrota

Continuemos analizando otros aspectos de la interacción argumentativa. Las preferencias pueden establecerse no sólo entre argumentos adversos sino también entre argumentos que apoyan una misma tesis. Pero cuando se da el primer caso estamos frente a un *rebatimiento* o *derrota*. Decimos que un argumento *arg1 rebate* o *derrota* a otro argumento *arg2* cuando *arg1* es incompatible con *arg2* (o algún subargumento *arg2'* de *arg2*) en el contexto, y *arg1* es preferido a *arg2* (resp. *arg2*'). Esta relación no es transitiva. Esto hace que las derrotas no sean suficientes para determinar la elección de los argumentos justificados ("ganadores") de un sistema. El problema se manifiesta en el hecho de que si un argumento *arg1* derrota a otro *arg2*, entonces *arg1* podría quedar justificado y *arg2* ser perdedor (fig. 1.a), pero si hay un tercer argumento *arg3* que derrota a *arg1*, entonces *arg3* y *arg2* resultarían justificados y *arg1* sería el perdedor (fig. 1.b). En efecto, la justificación es un proceso en el que todos los argumentos deben interactuar hasta agotar todas las instancias en que un argumento propuesto pueda ser derrotado.

Figura 2

2. Enfoques alternativos para la justificación de argumentos

Del mismo modo que para establecer un criterio de preferencias, el criterio para hallar los argumentos justificados puede depender del tipo de creencias representado. Un sistema puede confeccionarse para inferir información tentativa —qué nuevas creencias derivar de otras creencias— o para decidir un curso de acción —cómo actuar de acuerdo a las creencias.

Ejemplo 2. Supongamos que un agente debe diagnosticar una enfermedad ante la constatación de ciertos síntomas, a fin de sugerir la terapia conveniente. El agente sabe que bajo el síntoma *s* pueden manifestarse dos enfermedades distintas e_1 y e_2, a las cuales corresponde aplicar distintas terapias t_1 y t_2, respectivamente, mientras ambas terapias no deben ser seguidas a la vez. Esto puede expresarse del siguiente modo:

$\mathsf{K} = \{\neg(terapia(t_1) \wedge terapia(t_2))\}$
$\mathsf{R} = \{síntoma(s) \succ\!\!- enfermedad(e_1),\ síntoma(s) \succ\!\!- enfermedad(e_2),\ enfermedad(e_1) \succ\!\!- terapia(t_1),\ enfermedad(e_2) \succ\!\!- terapia(t_2)\}$

El agente construye los siguientes argumentos (obviamos la secuencia que los construye):

arg1: $\langle \{síntoma(s) \succ\!\!- enfermedad(e_1),\ enfermedad(e_1) \succ\!\!- terapia(t_1)\},\ \{síntoma(s)\},\ terapia(t_1)\rangle$

arg2: $\langle \{síntoma(s) \succ\!\!- enfermedad(e_2),\ enfermedad(e_2) \succ\!\!- terapia(t_2)\},\ \{síntoma(s)\},\ terapia(t_2)\rangle$

Si no hay criterio que permita decidir cuál de estos argumentos es el preferido, entonces no hay modo racional de *creer* cuál de las terapias es conveniente; por lo tanto, la indecisión es epistémicamente correcta. Sin embargo, si se procura decidir tomar un curso de acción siguiendo una de las terapias, entonces la indecisión resulta irracional desde el punto de vista *práctico*, siendo preferible tomar aleatoriamente cualquiera de las alternativas (ocurre lo mismo que en el famoso *problema del asno* de Buridan).

Si lo que deseamos es obtener las creencias justificadas más allá de los fines prácticos, entonces pueden resultar apropiadas las definiciones que determinan un único conjunto de argumentos justificados. Definen funciones que se las llama *asignaciones de status único*. Éstas permiten excluir todo argumento cuya justificación no esté decidida (ante la duda el agente se mantiene escéptico). En el ejemplo 2, una asignación de status único determinará un conjunto de argumentos justificados en el cual no se hallarán ni *arg1* ni *arg2*. Esto permite

representar el hecho de que el agente no se compromete *teóricamente* con ninguna de las dos terapias.

Si, en cambio, se procura obtener una decisión *práctica*, más allá de su justificación teórica, entonces pueden resultar más apropiadas las definiciones que determinan una relación que hace corresponder a un mismo contexto distintos conjuntos posibles de argumentos justificados. Éstas relaciones se denominan *asignaciones de status múltiples*. En el ejemplo anterior, una asignación de status múltiple determinará dos conjuntos posibles de argumentos justificados, uno conteniendo a *arg1* y el otro conteniendo a *arg2*, de modo que cualquier conjunto que se escoja determine la aplicación de una terapia. Este tipo de asignaciones suelen ser preferidas a las primeras, ya que permiten representar a la vez el compromiso teórico del agente con las justificaciones que obtiene, tomando como teóricamente aceptable sólo la intersección de todos los conjuntos posibles.

5. Conclusión

El estudio de los sistemas argumentativos ha permitido dilucidar distintos aspectos que comprende la argumentación rebatible. Si bien las soluciones halladas no pueden ser definitivas, puesto que el tratamiento de los distintos aspectos se lleva a cabo mediante el análisis de modelos a veces demasiado simplificados, al menos hemos avanzado en la comprensión de las estructuras más generales de la argumentación rebatible. Esto redunda no sólo en un enriquecimiento filosófico, sino también en haber hallado algunos buenos fundamentos que justifican la racionalidad del comportamiento de los programas de Inteligencia Artificial que implementan tales modelos de argumentación. Este éxito, relativo pero halagüeño, está inspirando nuevos enfoques de diversos temas ligados al razonamiento práctico, tales como la toma de decisiones, la teoría de juegos, la racionalidad económica y la negociación.

Referencias y bibliografía relevante

1. Bodanza, G. (2002) “Disjunctions, Contraposition and Specificity in Defeasible Argumentation”. *Logic Journal of The IGPL* 10 (1) (23-50).
2. 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 (321-357).
3. Kowalski, R. and F. Toni (1996) “Abstract Argumentation”. *Artificial Intelligence and Law* 4 (3-4) (275-396).
4. Kyburg, H. (1974) *The Logical Foundations of Statistical Inference*, Reidel, Dordrecht.
5. Lin, F. and Shoham, Y. (1989) “Argument Systems: A Uniform Basis For Nonmonotonic Reasoning”. *Proc. of the First International Conference on Principles of Knowledge Representation and Reasoning* (245-255), San Mateo, CA, Morgan Kaufmann Pub.
6. Loui, R. (1987) “Defeat Among Arguments: A System of Defeasible Inference”. *Computational Intelligence* 2 (100-106).
7. Parsons, S., Sierra, C. and Jennings, N. (1998) “Agents that Reason and Negotiate by Arguing”. Manuscrito. Aparecerá en *Journal of Logic and Computation*.
8. Pollock, J. (1990) *Nomic Probability And The Foundations Of Induction.* Oxford University Press.
9. Poole, D. (1985) “On the Comparison of Theories: Preferring the Most Specific

Explanation". *Proc. of the Ninth IJCAI*, Los Altos (144-147).
10. Poole, D. (1991) "The Effect of Knowledge on Belief: Conditioning, Specificity and the Lottery Paradox in Default Reasoning". *Artificial Intelligence* 49 (281-307).
11. Prakken, H. (1993) *Logical Tools For Modeling Legal Argument*. Tesis doctoral, Vrije Universiteit, Amsterdam.
12. Prakken, H. and G. Sartor (1996) "A Dialectical Model of Assessing Conflicting Arguments in Legal Reasoning". *Artificial Intelligence and Law* 4(3-4) (331-368).
13. Rescher, N. (1977) *Dialectics: A Controversy-Oriented Approach To The Theory Of Knowledge*, State University of New York Press, Albany.
14. Simari, G. and R. Loui (1992) "A Mathematical Treatment of Defeasible Reasoning". *Artificial Intelligence* 53 (125-157).
15. Tohmé, F. (1997) "Negotiation and Defeasible Reasons for Choice". *Proceedings of the AAAI Stanford Spring Symposium on Qualitative Preferences in Deliberation and Practical Reasoning* (95-102).
16. Verheij, B. (1996) "Two Approaches to Dialectical Argumentation: Admisible Sets and Argumentation Stages". Presentado en el *Computational Dialectics Workshop*, junio 3-7 1996, Bonn. Publicado como reporte SKBS/B3.A/96-01. Texto completo disponible en la World-Wide Web en http://nathan.gmd.de/projects/zeno/fapr/programme.html.
17. Vreeswijk, G. (1997) "Abstract Argumentation Systems". *Artificial Intelligence* 90 (225-279).