Explaining Best Decisions via Argumentation

Qiaoting Zhong[1], Xiuyi Fan[2], Francesca Toni[2], and Xudong Luo[1]

[1] Sun Yat-sen University, China
[2] Imperial College London, United Kingdom

Abstract. This paper presents an argumentation-based multi-attribute decision making model, where decisions made can be explained in natural language. More specifically, an explanation for a decision is obtained from a mapping between the given decision framework and an argumentation framework, such that best decisions correspond to admissible sets of arguments, and the explanation is generated automatically from dispute trees sanctioning the admissibility of arguments. We deploy a notion of rationality where best decisions meet most goals and exhibit fewest redundant attributes. We illustrate our method by a legal example, where decisions amount to past cases most similar to a given new, open case.

1 Introduction

Argumentation is often portrayed as a powerful method to support several aspects of decision-making [10, 18], especially when decisions need to be explained, but, except for few studies (notably [1, 9, 15]), it is not connected to means of formal evaluation sanctioning recommended decisions as rational. Moreover, existing approaches to argumentation-based decision making either lack automatic support for generating explanations (*e.g.*, [5]) or directly use the outputs of argumentation engines as explanations (*e.g.*, [12]), even though these may be obscure to the non-expert human users. To this end, this paper gives an argumentation-based decision-making model that can base the output of an argumentation engine to generate automatically argumentative explanations for rational decisions in natural language, so that humans can understand and trust the decisions recommended by our model.

In this paper, we deal with multi-attribute decision making problems where decisions may or may not fulfil goals depending on whether they exhibit attributes capable of reaching those goals. Our recommended decisions are rational in that they meet most goals (in the sense of [9]) but in addition *deviate minimally* by having as few redundant attributes as possible, not contributing to fulfilling the goals. Concretely, in the spirit of [9], we define a (provably correct) mapping between the kind of decision framework we consider and a specific Assumption-Based Argumentation (ABA) [6] framework, so that minimally deviating decisions correspond to admissible sets of arguments.

We use the argumentation mapping to generate explanations, in natural language, for explaining decisions, and in particular why some decisions are preferred to (more rational than) others. These natural language explanations are generated automatically, using an algorithm we design in this paper, from dispute trees computed as a standard argumentative explanation for decisions in ABA. Overall, for any pair of decisions, our

Index	No.	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	Sentence
1	245	0	1	0	1	0	1	0	0	0	1y+$1,000
2	97	1	0	0	0	0	0	1	0	0	10y+3y+$10,000
3	420	1	0	0	1	1	0	0	1	0	6m+$1,000
4	96	1	0	0	0	0	1	0	1	0	3y+$3,000
5	48	1	0	0	0	1	0	0	1	0	6m+$1,000
6	751	1	0	0	0	0	1	0	1	0	5y+$5,000
7	412	1	0	0	1	0	0	0	0	0	5m+$1,000
8	1962	1	0	0	0	0	0	1	0	1	7y+$7,000
9	389	1	0	0	1	1	0	0	0	0	4m+$1,000
10	686	1	0	1	0	1	0	0	1	0	6m+$1,000
11	355	1	0	0	0	0	0	1	1	0	10y+3y+$10,000

Table 1. A fragment of the Past Cases Characteristics data, where $a_1, \cdots, a_9$ stand for "older than 18", "age between 16 and 18", "burglary", "repeatedly", (value of goods) "large amount", "huge amount", "extremely huge amount", "goods found" and "accessory", respectively.

technique can give succinct, human understandable descriptions to explain why one decision is better than another, which is particularly important for applications.

Throughout we motivate and illustrate our work with the following legal example.

Example 1. In the practice of law, when lawyers, judges, jury members or other legal entities receive a new case, they need to identify and compare similar past cases, and then use the (court) sentence for the past cases as a prediction for the possible sentence for the open case [13]. This can be viewed as a decision making problem, *i.e.*, past cases are alternative decisions whose rationality can be measured once the sentence of the new case is out: the closer the sentence to the ones for the past cases, the more rational the choice of these past cases as similar. Consider the 11 cases summarised in Table 1,[3] where: each case has a number of attributes, *e.g.*, "the item is of large value" (a_5 in Table 1) or "the goods have been found" (a_8 in Table 1), and a sentence, *e.g.*, "1 year of imprisonment with $1,000 fine" or "10 years of imprisonment, 3 years deprive of political right with $10,000 fine". For each case, attributes can have value 1 (if the case has that attribute) or 0 (if it does not). For instance, the row for the first case in Table 1 represents:

> *The defendant in case No. 245, with age between 16 and 18 (a_2), stole repeatedly (a_4), and the value of the stolen goods was huge (a_6). The resulting sentence was 1 year imprisonment with $1,000 fine.*

Throughout the paper we view case No. 355 (case 11 in Table 1) as new (thus ignoring its sentence). Case No. 97 is the past case closer to case No. 355 in terms of the final sentence, and thus it can be deemed the most similar. Case No. 97 is the case matching most attributes of case No. 355 while also "deviating minimally" from it (case No. 97 only deviates from case No. 355 on a_8). We will define rational decisions following this intuition, and give an argumentative counterpart thereof that can be used to explain why other past cases are not so similar as case No. 97 to case No. 355.

[3] These cases, all concerning theft, are adapted from real cases from the Nanhai District People's Court in the city of Foshan, Guangdong Province, China.

The rest of this paper is organised as follows. Section 2 recalls necessary background knowledge. Section 3 defines minimally deviating decisions. Section 4 describes the argumentative counterpart of rational decisions, defined in terms of minimal deviation. Section 5 discusses how to compare decisions. Section 6 designs the algorithm for the generation of natural language explanation for (rational) decisions. Section 7 discusses related work. Finally, Section 8 concludes this paper with future work.

2 Background

This work relies upon Assumption-Based Argumentation (ABA) [6] and the decision frameworks of [9]. We recap them in this section.

An *ABA framework* is a tuple $\langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \mathcal{C} \rangle$, where $\langle \mathcal{L}, \mathcal{R} \rangle$ is a deductive system, with *language* $\mathcal{L}$ and rule set $\mathcal{R} = \{s_0 \leftarrow s_1, \ldots, s_m \mid s_0, \cdots, s_m \in \mathcal{L}\}$; $\mathcal{A} \subseteq \mathcal{L}$ is a (non-empty) set, referred to as the *assumptions*; and $\mathcal{C}$ is a total mapping from $\mathcal{A}$ into $2^{\mathcal{L}} \setminus \{\{\}\}$, *i.e.*, $\mathcal{C}(\alpha)$ is the *contrary* of $\alpha \in \mathcal{A}$. Given *rule* $\rho = s_0 \leftarrow s_1, \ldots, s_m$, s_0 is the *head* (denoted $Head(\rho) = s_0$) and $s_1, \ldots, s_m$ constitute the *body* (denoted $Body(\rho) = \{s_1, \ldots, s_m\}$). If $m = 0$, ρ is represented as of $s_0 \leftarrow$ and $Body(\rho) = \{\}$.

In ABA, *arguments* are deductions of claims using rules and supported by assumptions, and *attacks* are directed at assumptions:

- *an argument for claim $c \in \mathcal{L}$ supported by $S \subseteq \mathcal{A}$* ($S \vdash c$ for short) is a (finite) tree with nodes labelled by sentences in $\mathcal{L}$ or by the symbol τ,[4] such that the root is labelled by c, leaves are either τ or assumptions in S, and a non-leave s has as many children as the elements in the body of a rule with head s, in an one-to-one correspondence with the elements of this body; and
- $S_1 \vdash c_1$ *attacks* $S_2 \vdash c_2$ iff $c_1 \in \mathcal{C}(\alpha)$ for some $\alpha \in S_2$.

A set of arguments is *admissible* if and only if it does not attack any argument it contains but attacks all arguments attacking it.[5] Admissible sets of arguments can be characterised in terms of *dispute trees* [7], namely trees with *proponent (P)* and *opponent (O)* nodes, labelled by arguments, and such that arguments labelling a node attack the argument in their parent node. Each P-node has all their attacking arguments as its children, and each O-node has one child only. If no arguments in a dispute tree label a P-node as well as an O-node, then the dispute tree is *admissible* and the set of all arguments labelling P-nodes (called *defence set*) is admissible [7]. Argumentation engines, such as `proxdd`,[6] compute admissible dispute trees and admissible sets of arguments.

A *decision framework* is a tuple $\langle \mathtt{D}, \mathtt{A}, \mathtt{G}, \mathtt{DA}, \mathtt{GA} \rangle$, with a set of decisions $\mathtt{D} = \{d_1, \cdots, d_n\}$, $n > 0$; a set of attributes $\mathtt{A} = \{a_1, \cdots, a_m\}$, $m > 0$; a set of goals $\mathtt{G} = \{g_1, \cdots, g_l\}$, $l > 0$; two tables, DA (of size $n \times m$) and GA (of size $l \times m$),[7] such that

- every $\mathtt{DA}_{i,j} (1 \leq i \leq n, 1 \leq j \leq m)$ is either 1, representing that alternative decision d_i *has* attributes a_j, or 0, otherwise; and

[4] $\tau \notin \mathcal{L}$ stands for "true" and is used to represent the empty body of rules [6].

[5] An argument set As attacks an argument B iff some $A \in As$ attacks B, and an argument A attacks an argument set Bs iff A attacks some $B \in Bs$.

[6] `www.doc.ic.ac.uk/~rac101/proarg`

[7] We use $\mathtt{X}_{i,j}$ to represent the cell in row i and column j in $\mathtt{X} \in \{\mathtt{DA}, \mathtt{GA}\}$.

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
g_1	1	0	0	0	0	0	0	0	0
g_2	0	0	0	0	0	0	1	0	0
g_3	0	0	0	0	0	0	0	1	0

Table 2. GA table for Example 2.

– every $\texttt{GA}_{i,j} (1 \leq i \leq l, 1 \leq j \leq m)$ is either 1, representing that goal g_i is *satisfied* by attribute a_j, or 0, otherwise.

The column orders in both DA and GA are the same, and the *indices* of decisions, goals, and attributes in DA and GA are the row numbers of the decisions and goals and the column number of attributes in DA and GA, respectively. $\mathcal{DEC}$ and $\mathcal{DF}$ denote the set of all possible decisions and the set of all possible decision frameworks, respectively.

A decision $d \in \texttt{D}$ with row index i in DA *meets* a goal $g \in \texttt{G}$ with row index j in GA if and only if there is an attribute $a \in \texttt{A}$ with column index k in both DA and GA, such that $\texttt{DA}_{i,k} = 1$ and $\texttt{GA}_{j,k} = 1$. $\gamma(d) \subseteq \texttt{G}$ denotes the set of goals met by d.

A mapping $\psi : \mathcal{DF} \mapsto 2^{\mathcal{DEC}}$ is a *decision function* if, for $df = \langle \texttt{D}, \texttt{A}, \texttt{G}, \texttt{DA}, \texttt{GA} \rangle$, $\psi(df) \subseteq \texttt{D}$. $\psi(df)$ is the set of decisions that are selected with respect to ψ. Ψ denotes the set of all decision functions.

Strongly dominant decisions meet all goals. *Weakly dominant* decisions meet goals that are not met by other decisions. Formally, given $df = \langle \texttt{D}, \texttt{A}, \texttt{G}, \texttt{DA}, \texttt{GA} \rangle$,

– $\psi_s \in \Psi$ is a *strongly dominant* decision function iff $\forall d \in \psi_s(df)$, $\gamma(d) = \texttt{G}$.
– $\psi_w \in \Psi$ is a *weakly dominant* decision function iff $\forall d \in \psi_w(df)$, $\nexists d' \in \texttt{D} \backslash \{d\}$ such that $\gamma(d) \subset \gamma(d')$.

In the remainder of this paper, unless otherwise specified, we will assume a generic decision framework $df = \langle \texttt{D}, \texttt{A}, \texttt{G}, \texttt{DA}, \texttt{GA} \rangle \in \mathcal{DF}$.

3 Minimally Deviating Decisions

This section discusses decision criteria. Meeting goals is crucial in rational decision selection, but it does not always allow to discriminate amongst decisions, and more importantly it ignores other factors in decision making: the presence of "redundant" attributes, illustrated below.

Example 2. (Example 1 continued.) With respect to the new case No. 355, the decision problem can be formalised as df in which $\texttt{D} = \{d_1, \cdots, d_{10}\}$, $\texttt{A} = \{a_1, \cdots, a_9\}$, $\texttt{G} = \{g_1, g_2, g_3\}$ (where g_1, g_2 and g_3 stand for "older than 18", "extremely huge amount" and "goods found", respectively), DA is adapted from Table 1 (without case No. 355 and sentences) and GA is in Table 2. Then, using the decision function ψ_s, no strongly dominant decisions exist. Since both d_2 (case No. 97) and d_8 (case No. 1962) meet two goals, *i.e.*, g_1 and g_2, they are both weakly dominant and thus equally good according to the decision function ψ_w. However, these two cases have different sentences (as shown in Table 1). But case No. 1962 also has attribute a_9 that contributes to no goals and makes case No. 1962 distinct from our new case. Thus, we can deem case No. 97 more similar to the new one. This is legitimated by the actual sentence for case No. 355, since this is the same as the sentence for case No. 97.

Thus, we need new decision criteria. Intuitively, given a decision framework, a decision d is optimal iff d is (strongly or weakly) dominant with the fewest redundant attributes. Formally, we have:

Definition 1. *Let $\alpha \in \mathtt{A}$ and i be the column index in* DA *and* GA*. Then α is a* deviating attribute *iff $\forall g \in \mathtt{G}$, if g has row index j in* GA*, then $\mathtt{GA}_{j,i} \neq 1$. The set of* deviating attributes decision d has *is denoted as $\lambda(d)$.*

For our legal example, since $\mathtt{GA}_{1,9} = 0$, $\mathtt{GA}_{2,9} = 0$ and $\mathtt{GA}_{3,9} = 0$, a_9 is a deviating attribute. Similarly, $a_2, \cdots, a_6$ are deviating. Then, $\lambda(d_1) = \{a_2, a_4, a_6\}$, $\lambda(d_2) = \{\}$ and so on. Intuitively, $\alpha \in \lambda(d)$ means that d has α but α fulfils no goals in G.

Intuitively, *minimally-deviating* decisions have a minimal number of deviating attributes with respect to set inclusion. They are the output of minimally-deviating decision functions. Formally, we have:

Definition 2. *$\psi_m \in \Psi$ is a* minimally-deviating *decision function iff $\forall d \in \psi_m(df)$, $\nexists d' \in \mathtt{D}$ such that $\lambda(d') \subset \lambda(d)$.*

In words, a decision d is minimally-deviating if and only if there does not exist d' such that the set of deviating attributes that d' has is a proper subset of those that d has. For our legal example, it is easy to see that d_2 is minimally-deviating.

Minimal deviation is about decision having attributes, hence it is orthogonal to *dominance* (see Section 2), concerning decisions meeting goals. Thus, we can select decisions in a two-step process: first find dominant decisions, and then, amongst these, further select the minimally deviating decisions. Thus, we introduce *sub-frameworks* to refine decisions on grounds of deviation as follows:

Definition 3. *Given $\mathtt{D}' \subseteq \mathtt{D}$, the* sub-framework *of df w.r.t. $\mathtt{D}'$ is a decision framework $\langle \mathtt{D}', \mathtt{A}, \mathtt{G}, \mathtt{DA}', \mathtt{GA} \rangle$ such that $\mathtt{DA}'$ is the restriction of* DA *that contains only rows for d_i such that $d_i \in \mathtt{D}'$.*

We can then combine strongly/weakly dominance (s-dominance/w-dominance for short) and minimal deviation as follows:

Definition 4. *Let df_s be the sub-framework of df w.r.t. $\psi_s(df)$, and df_w be the sub-framework of df w.r.t. $\psi_w(df)$. Then: (i) $\psi_{ms} \in \Psi$ is a* Minimally-Deviating S-Dominant (MDSD) *decision function if $\forall d \in \mathtt{D}$, $d \in \psi_{ms}(df)$ iff $d \in \psi_m(df_s)$; and (ii) $\psi_{mw} \in \Psi$ is a* Minimally-Deviating W-Dominant (MDWD) *decision function if $\forall d \in \mathtt{D}$, $d \in \psi_{mw}(df)$ iff $d \in \psi_m(df_w)$.*

For our legal example, d_2 (case No. 97) is MDWD as it is the only weakly dominant decision that has no deviating attributes.

The following result relates MDSD and MDWD decisions.

Proposition 1. *Let $D_s = \psi_s(df)$, $D_w = \psi_w(df)$, $D_{ms} = \psi_{ms}(df)$ and $D_{mw} = \psi_{mw}(df)$. If $D_s \neq \{\}$ then $D_{ms} = D_{mw}$.*

Proof. Let df_s be the sub-framework of df w.r.t. D_s, and df_w be the sub-framework of df w.r.t. D_w. We firstly prove that $D_{mw} \subseteq D_{ms}$. For any $d \in D_{mw}$, we have $d \in \psi_m(df_w)$ by Definition 4. Since $D_s \neq \{\}$, we have $D_s = D_w$ (by Proposition 4 in [9]). Thus, we have $df_w = df_s$. Therefore, $d \in \psi_m(df_s)$ holds. By Definition 4, we have $d \in D_{ms}$. Similarly, we can prove that $D_{ms} \subseteq D_{mw}$. □

4 Argumentative Counterpart of Minimally Deviating Decisions

After introducing a new decision criterion, *minimal deviation* and related concepts of MDSD and MDWD decisions, this section presents a method to map the problem of determining minimally deviating decisions onto the problem of finding admissible sets of argument in an argumentation framework. This mapping serves as a means to generate dispute trees that explain decisions in its own right and will be used in Section 6 to feed into the algorithm for providing natural language explanations.

Concretely, we use ABA to develop the mapping. The idea is that for a decision framework and a given decision function (*e.g.*, ψ_{ms} or ψ_{mw}), an *equivalent* ABA framework with rules, assumptions and contraries is constructed; then the computation of selected decisions can be performed within the ABA framework via standard argumentation semantics computation, using tools such as `proxdd` (see Section 2). We show that such a mapping is not only sound and complete but also able to generate an argumentative explanation.

We start with giving the ABA construction for MDSD decisions.

Definition 5. *Let df* $= \langle \mathtt{D}, \mathtt{A}, \mathtt{G}, \mathtt{DA}, \mathtt{GA} \rangle$ *be such that* $|\mathtt{D}| = n$, $|\mathtt{A}| = m$ *and* $|\mathtt{G}| = l$. *The* MDSD ABA framework *corresponding to* $\langle \mathtt{D}, \mathtt{A}, \mathtt{G}, \mathtt{DA}, \mathtt{GA} \rangle$ *is* $\langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \mathcal{C} \rangle$, *where:*

- $\mathcal{L}$ *is the language.*
- $\mathcal{R}$ *consists of the following rules:*[8]

$\forall d_k \in \mathtt{D}, g_j \in \mathtt{G}, a_i \in \mathtt{A}$: *isD*$(d_k) \leftarrow$; *isG*$(gj) \leftarrow$; *isA*$(a_i) \leftarrow$; (1)

$\forall 1 \leq k \leq n, 1 \leq i \leq m$, if $\mathtt{DA}_{k,i} = 1$ then *hasAttr*$(d_k, a_i) \leftarrow$; (2)

$\forall 1 \leq j \leq l, 1 \leq i \leq m$, if $\mathtt{GA}_{j,i} = 1$ then *satBy*$(g_j, a_i) \leftarrow$; (3)

de$(X, Y) \leftarrow$ *isD*(X), *isA*(Y), *hasAttr*(X, Y), *deviate*(Y); (4)

notDeviate$(Y) \leftarrow$ *satBy*(X, Y), *isG*(X), *isA*(Y); (5)

notMSD$(X) \leftarrow$ *sd*(X'), *worse*(X, X'), *nWorse*(X', X), *isD*(X), *isD*(X'); (6)

nSD$(X) \leftarrow$ *nMet*(X, Z), *isD*(X), *isG*(Z); (7)

met$(X, Y) \leftarrow$ *hasAttr*(X, Z), *satBy*(Y, Z), *isD*(X), *isG*(Y), *isA*(Z); (8)

worse$(X', X) \leftarrow$ *isD*(X), *isD*(X'), *isA*(Y), *de*(X', Y), *nAttr*(X, Y). (9)

- $\mathcal{A} = \{\{$*nWorse*$(d_r, d_k) | d_k, d_r \in \mathtt{D}\} \cup \{$*nMet*$(d_k, g_j) \mid d_k \in \mathtt{D}, g_j \in \mathtt{G}\} \cup$
  $\{$*ms*(d_k), *sd*(d_k), *deviate*(a_i), *nAttr*$(d_k, a_i) \mid d_k \in \mathtt{D}$, $a_i \in \mathtt{A}\}\}$. (10)
- $\mathcal{C}($*ms*$(d_k)) = \{$*notMSD*(d_k), *nSD*$(d_k)\}$; $\mathcal{C}($*sd*$(d_k)) = \{$*nSD*$(d_k)\}$; (11)

  $\mathcal{C}($*deviate*$(a_i)) = \{$*notDeviate*$(a_i)\}$; $\mathcal{C}($*nAttr*$(d_k, a_i)) = \{$*hasAttr*$(d_k, a_i)\}$; (12)

  $\mathcal{C}($*nWorse*$(d_r, d_k)) = \{$*worse*$(d_r, d_k)\}$; $\mathcal{C}($*nMet*$(d_k, g_j)) = \{$*met*$(d_k, g_j)\}$. (13)

We explain some crucial aspects of the above definition as follows. A decision d_k is assumed to be MDSD by declaring it as an assumption $ms(d_k)$ (see (10)). d_k is not MDSD under either of the two conditions: (i) d_k is not strongly dominant, or (ii) d_k is not minimally-deviating. Hence, the contraries of $ms(d_k)$ are $nSD(d_k)$ and $notMSD(d_k)$, respectively (see (11)). For any decision X, it is not minimally-deviating ($notMSD(X)$) if there exists some decision X' such that X' is strongly dominant ($sd(X')$) and X contains some deviating attribute which X' does not ($worse(X,$

[8] We use *schemata* with variables to represent compactly all rules that can be obtained by instantiating the variables over the appropriate domains.

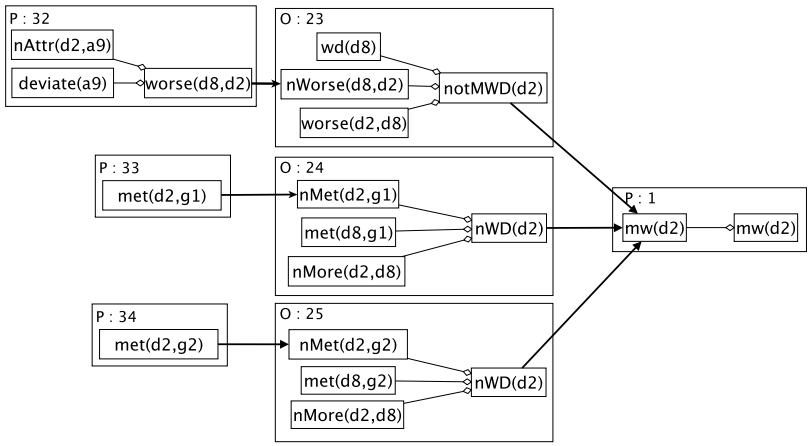

Fig. 1. A fragment of the admissible dispute tree for our legal example. Individual arguments are wrapped in boxes, the claim of an argument is the inner box on the right and assumptions and facts (rules with empty bodies) are the inner boxes on the left. Arguments are P or O (see Section 2). If there is only one inner box in an argument, then this is supported by the empty set (of assumptions). Attacks are labelled as arrows between outer boxes (the argument).

X')); and X' does not contain any more deviating attributes than X ($nWorse(X', X)$) (see (6)). A decision X is not strongly dominant ($nSD(X)$) if there exists some goal Z that X does not meet ($nMet(X, Z)$) (see (7)).

This mapping onto ABA for MDSD decisions can serve as the basis for the computation for MDSD decisions by virtue of the following theorem:

Theorem 1. *Let AF be the MDSD ABA framework corresponding to df. Then $\forall d_k \in \mathtt{D}$, $d_k \in \psi_{ms}(df)$ iff $\{ms(d_k)\} \vdash ms(d_k)$ belongs to an admissible set of arguments in AF.*

This theorem can be drawn from Definitions 4 and 5 as well as the definition of admissibility in ABA (see Section 2). For the sake of space, the proof is omitted.

Definition 6. *Let $\langle \mathcal{L}, \mathcal{R}, \mathcal{A}, \mathcal{C} \rangle$ be the MDSD ABA framework corresponding to df. The* MDWD ABA framework *corresponding to df is $\langle \mathcal{L}, \mathcal{R}', \mathcal{A}', \mathcal{C}' \rangle$, where:*

- *$\mathcal{R}'$ is derived from $\mathcal{R}$ by replacing rules with heads notMSD(X) and nSD(X) with the rules:*
  notMWD$(X) \leftarrow$ wd(X'), worse(X, X'), nWorse(X', X), isD(X), isD(X'),
  nWD$(X) \leftarrow$ met(Y, Z), nMet(X, Z), nMore(X, Y), isD(X), isD(Y), isG(Z),
  nMore$(X, Y) \leftarrow$ met(X, Z), nMet(Y, Z), isG(Z), isD(X), isD(Y);
- *$\mathcal{A}'$ is $\mathcal{A} \cup \{$nMore$(d_r, d_k) | d_r, d_k \in \mathtt{D}\}$, where ms$(d_k)$ and sd(d_k) are replaced by mw(d_k) and wd(d_k), respectively.*
- *$\mathcal{C}'($nMore$(d_r, d_k)) = \{$more$(d_r, d_k)\}$ and otherwise $\mathcal{C}'$ is $\mathcal{C}$ except that ms(d_k), nSD(d_k), notMSD(d_k) and sd(d_k) are replaced by mw(d_k), nWD, notMWD(d_k) and wd(d_k), respectively.*

Similarly, ABA can also be used for MDWD by virtue of the following theorem:

Theorem 2. *Let F be the MDWD ABA framework corresponding to df. Then $\forall d_k \in \mathtt{D}$, $d_k \in \psi_{mw}(df)$ iff $\{mw(d_k)\} \vdash mw(d_k)$ belongs to an admissible set of arguments in F.*

This theorem can be drawn from Definitions 4 and 6 as well as the definition of admissibility in ABA (see Section 2). The proof details are omitted for lack of space.

We illustrate the computation of selected decisions in ABA for our legal example:

Example 3. (Example 2 continued.) Given the MDSD ABA framework corresponding to our legal *df* (see Definition 5), $\{\text{mw}(d_2)\} \vdash \text{mw}(d_2)$ belongs to an admissible set, as shown by the fragment of an admissible dispute tree in Figure 1, adapted from the output of `proxdd`. This tree illustrates the argumentative explanation aspect of ABA computation. The root argument of the tree (right-most box labelled P:1) claims that d_2 is MDWD. This claim is attacked by three different arguments, boxes O:23 - O:25, giving reasons for why d_2 is not MDWD as follows:

- (O:23) d_8 is better than d_2;
- (O:24) d_2 is not weakly dominant since d_2 does not meet goals that are not met by d_8, but d_8 meets the goal g_1 that are not met by d_2; and
- (O:25) d_2 is not weakly dominant since d_2 does not meet goals that are not met by d_8, but d_8 meets the goal g_2 that are not met by d_2.

These three arguments are counter-attacked by arguments in P:32 - P:34, respectively:

- (P:32) d_8 has deviating attribute a_9 but d_2 does not, so d_8 is worse than d_2; and
- (P:33) / (P:34) d_2 meets the goal g_1 / g_2 (respectively).

5 Comparing Decisions

We have considered several criteria for evaluating decisions, *i.e.*, strong/weak dominance, MDSD and MDWD. Here we will formally compare decisions in different categories and explain their differences. That is, we give a mechanism to rank decisions that meet different criteria. We start with defining the *better-than* relation between decisions:

Definition 7. *For all $d, d' \in \mathtt{D}$, d is* better than *d', denoted $d \succ d'$, iff:*

(i) $d \in \psi_s(df)$ *and* $d' \notin \psi_s(df)$, *or*
(ii) $d \in \psi_w(df)$ *and* $d' \notin \psi_w(df)$, *or*
(iii) $d \in \psi_{ms}(df)$ *and* $d' \in \psi_s(df)$ *but* $d' \notin \psi_{ms}(df)$, *or*
(iv) $d \in \psi_{mw}(df)$ *and* $d' \in \psi_w(df)$ *but* $d' \notin \psi_{mw}(df)$.

d is as good as *d', denoted $d \sim d'$, iff neither $d \succ d'$ nor $d' \succ d$.*

The intuition behind Definition 7 is that: (1) minimally deviating strongly/weakly dominant decisions are better than strongly/weakly dominant decisions, which in turn are better than non-strongly/non-weakly dominant decisions; and (2) if it is not the case that one decision is better than the other, then they are equally good.

Note that if the set of strongly dominant decisions D is not empty, then D is also weakly dominant (by Proposition 4 in [9]). Hence, strongly dominant decisions are as good as weakly dominant decisions.

For our legal example, since d_2 (case No. 97) is the only MSWD and d_8 (case No. 1962) is weakly dominant, we have $d_2 \succ d_8$ according to Definition 7(iv).

To ensure that our notions of "better than" and "as good as" are well defined, we need to show that for any two decisions in our decision framework, they can always be compared using our notions. The following several results sanction this:

Proposition 2. *For all d_1, d_2, $d_3 \in \mathtt{D}$, if $d_3 \succ d_2$ and $d_2 \succ d_1$, then $d_3 \succ d_1$.*

Proposition 3. *$\sim$ is an equivalence relation on $\mathtt{D}$.*

Propositions 2 and 3 can be proved by using mathematical induction and contradiction, respectively. For the sake of space, their details are omitted here.

Given the quotient set $\mathtt{D}/\sim$ of all equivalence classes, we define a binary relation $\geq$ on $\mathtt{D}/\sim$ as follows:

Definition 8. *Given $d \in \mathtt{D}$, let $[d]$ denote the equivalence class to which d belongs. For any $[d], [d'] \in \mathtt{D}/\sim$, $[d] \geq [d']$ iff $d \succ d'$ or $d \sim d'$.*

Proposition 4. *$\geq$ is a total order on $\mathtt{D}/\sim$.*

Proof. (Sketch) We need to prove that $\geq$ is anti-symmetric, transitive and total. (i) Anti-symmetry. If $[d_1] \geq [d_2]$, then $d_1 \succ d_2$ or $d_1 \sim d_2$. For the former, $d_2 \not\succ d_1$. Since $[d_2] \geq [d_1]$, we have $d_2 \sim d_1$, which leads to $[d_1] = [d_2]$. For the latter, $[d_1] = [d_2]$ certainly. (ii) Transitivity. Since $\succ$ and $\sim$ are proven as transitive, it is easy to see that $\forall d_1, d_2, d_3 \in \mathtt{D}$, if $d_1 \sim d_2$ and $d_2 \succ d_3$, then $d_1 \succ d_3$; and if $d_1 \succ d_2$ and $d_2 \sim d_3$, then $d_1 \succ d_3$. (iii) Totality can be proven by contradiction. □

Proposition 4 actually says that any two decisions in a decision framework can be compared with respect to the "better than" or "as good as" relation given in Definition 7.

6 Natural Language Explanation

Dispute trees give a comprehensive explanation to decisions. However, they are less useful for comparing decisions, *i.e.*, for answering the question: *Why is d_i better than d_j?* Transparently answering such a comparison question is crucial for human, *e.g.*, to trust the decision that is recommended by computer systems. Hence, in this section we will design an algorithm to extract information from dispute trees and then generate natural language explanations specifically answering pairwise comparison questions. This algorithm corresponds to the decision ranking mechanism described in Section 5.

Figures 2-4 show our algorithm for generating natural language texts explaining the rationale in the decision making, where Figures 3 and 4 give subroutines for procedures used in Figure 2. We illustrate these algorithms with the legal example.

Example 4 (Example 3: continued). From earlier discussions, we know that d_2 is the best decision. Comparing d_2 with d_8, the algorithm in Figure 2 provides an explanation. Since argument $\{mw(d_2)\} \vdash mw(d_2)$ is admissible as d_2 is MDWD, the algorithm compares d_2 and d_8 with the dispute tree shown in Figure 1. Since argument $\{mw(d_8)\} \vdash mw(d_8)$ is not admissible as d_8 is not MDWD, and *nWD*(d_8) is not the conclusion of any P argument (indicating that d_8 is not weakly dominant), it will check whether or not *worse*(d_8, d_2) (d_8 is worse than d_2 for having more deviating attributes) is the claim of any P argument. Since *worse*(d_8, d_2) is indeed the claim of argument P: 32, which contains $deviate(a_9)$, indicating that d_8 has more deviating attributes (a_9) than d_2, the system outputs:

> *Both d_2 and d_8 meet most goals, however, d_2 is better than d_8 as d_2 has fewer deviating attributes. For example, d_8 has deviating attribute a_9 but d_2 does not.*

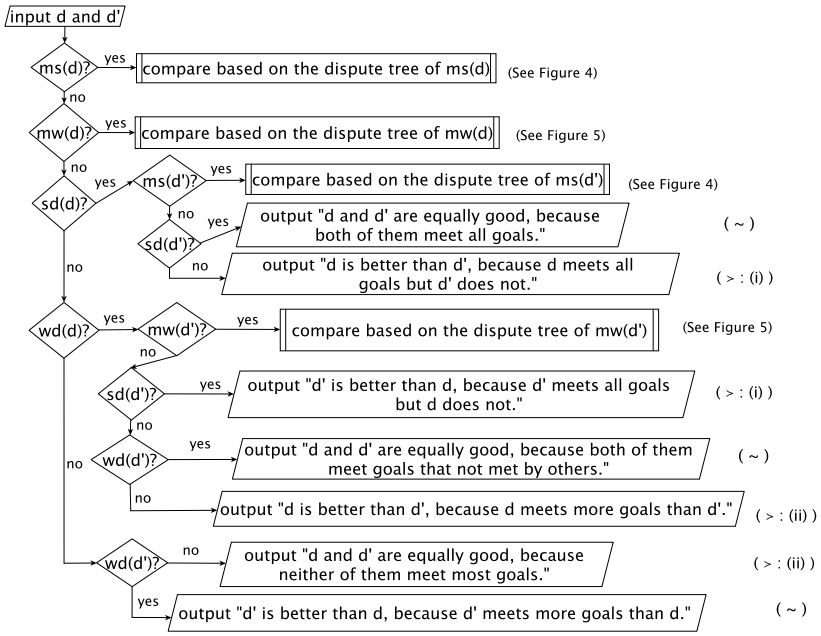

Fig. 2. Algorithm for automated explanation (labels on the right are used in proofs).

Theorem 3. *The algorithm shown in Figures 2, 3 and 4 is sound and complete w.r.t. the ordering given in Definition 7.*

Proof. (Sketch) With the labels beside each output in Figures 2-4, we can easily prove completeness. For soundness, here we just consider the case of $d \succ d'$ since $d \in \psi_s(df)$ and $d' \notin \psi_s(df)$, others can be proved similarly. Since $d \in \psi_s(df)$, if $ms(d)$ is admissible, according to Figure 3, the output is "d is better than d' as d meets all goals but d' does not"; otherwise, by Proposition 1, we have $d \notin \psi_{mw}(df)$. Then from Figure 2, the output is "d is better than d' because d meets all goals but d' does not". □

Putting everything together, based on the corresponding ABA framework, we see that case No. 97 is the most similar case, because only it is MDWD. Case No. 97 fully agrees with the sentence for case No. 355 (*i.e.*, 10 years of imprisonment, 3 years deprive of political right with $10,000 fine), so indeed it can be deemed to be the most similar, thus validating MDWD as a suitable decision criterion here. In addition, when being posed with the question: "*why is case No. 97 better than case No. 245?*", our algorithm gives:

> *No. 97 is better than No. 245 because No. 97 meets more goals. For example, No. 245 does not meet goal "the defendant is older than 18" but No. 97 does.*

7 Related work

Some studies exist about argumentation-based decision making and analysis. For example, in [9], Fan *et al.* give basic decision frameworks and functions, and in [8] they discuss how preferences over goals can be expressed and incorporated. Our work has

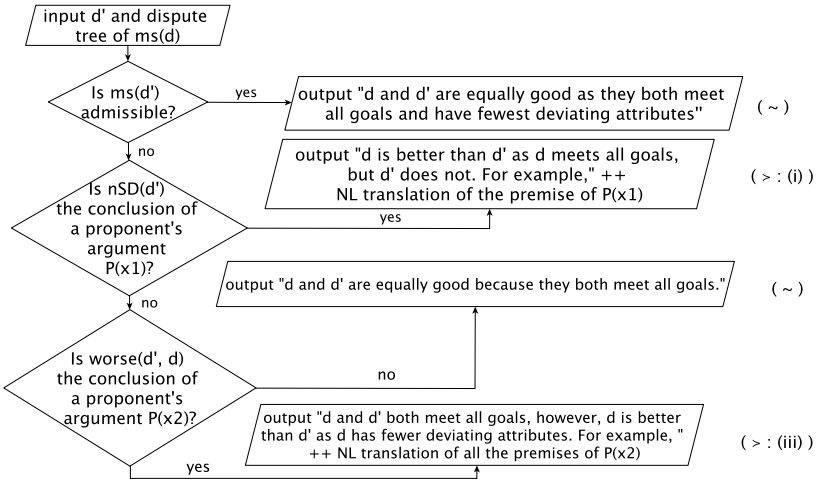

Fig. 3. Comparison based on the dispute tree of ms(d) (labels on the right are used in proofs).

extended theirs by introducing new decision criteria, a decision comparison method and the algorithms for generating natural language texts to explain the selected decisions. Müller and Hunter [17] present an argumentation-based system for decision analysis. Their work is based on ASPIC+ [20], whereas our work is based on ABA. Moreover, they present a method for generating decisions, whereas we present an algorithm for the generation of natural language explanations for decisions.

Natural Language Generation (NLG) is the natural language processing task of producing readable text in ordinary language, usually from complex data streams. A wide range of practical uses of NLG has been studied, including writing weather forecasts [22], summarising medical data [19] or generating hypothesis [21]. To the best of our knowledge, ours is the first NLG work for argumentation-based decision making.

Heras *et al.* [11] have proposed a case-based argumentation framework for agent societies for customer support. By being engaged in argumentation dialogues, agents reach agreements for best solutions. The differences between their work and ours are: (i) they use value-based argumentation [18], whereas we use ABA; and (ii) they justify the reasoning process by showing dialogue graphs, while we generate natural language text, which is (arguably) easier for ordinary human users to understand.

Our model can be used for case-based reasoning (CBR), but it is quite different from existing work. For example, though both HYPO [3, 4] and our method can identify the most similar cases, only ours can give an argumentative natural-language explanation ranking. Moreover, our work is actually a generic decision framework that can be used in any other domains, while HYPO is specific for law. Moreover, Armengol and Plaza [2] provide an explanation scheme for similarities in CBR, through which the users can understand why some cases are retrieved, but cannot know why others are not. Our work can do both by: (i) argumentative explanation with ABA computation and (ii) our natural language explanation for pairwise decision comparison. In addition, McSherry [16] applies a CBR approach to a recommender system that offers benefits in explaining the recommendation process, but its focus is on the efficiency and trans-

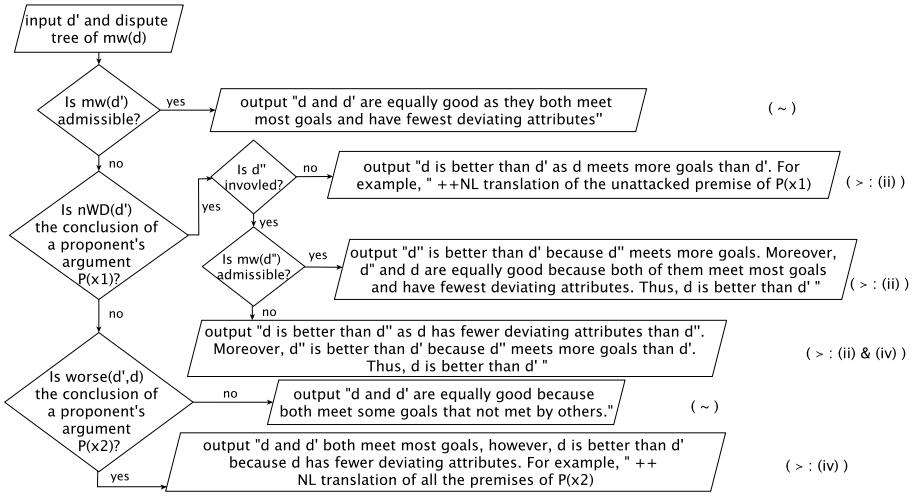

Fig. 4. Comparison based on the dispute tree of mw(d) (labels on the right are used in proofs).

parency of the recommendation process by explaining relevant questions being asked to users. Instead, our work focuses on multi-attribute decision making and the natural language explanation for a decision that our model recommend.

Finally, our model can be viewed as a general framework for the explanation of multi-attribute decision models. Similarly, in [14], a selected decision can be explained by the analysis of the values of weights on criteria together with the overall scores of the decisions. However, we do not consider weighted-based decision model but rather the natural language generation of pairwise comparison explanation.

8 Conclusion

This paper presents a new decision criterion, *minimal deviation*, and its combination with two notions of *dominance* to select decisions that meet more goals but with fewer "redundant" attributes (*i.e.*, attributes not contributing to meeting goals). We have developed mappings onto ABA for the two resulting types of decisions, not only serving as a basis for decision selection but also contributing to better explaining the selection.

Moreover, we have formally defined a decision ranking mechanism by giving a total ordering amongst all decisions. The argumentative counterpart used to identify best decisions also serves as foundation to provide natural language explanations for why one decision is better than another. Our natural language explanation algorithm fully takes advantage of the potential of argumentative decision making to support transparent explanations automatically. Finally, we have illustrated our method in the legal domain, for identifying the most similar cases in a repository of past cases, and more importantly for explaining, in natural language, why a case is more similar than others, which is critical for lawyers to trust the case(s) that our model recommends.

In the future, we plan to expand our legal case-study to larger repositories. It would also be interesting to consider decision criteria involving both attribute deviation and preference ranking over goals. It is also worth furthering the link between argumentation-

based decision making and other decision making methods, and applying our framework to other application domains. In particular, this will allows us to ascertain the generality of our method.

Acknowledgements

This research was supported by the EPSRC TRaDAr project: EP/J020915/1, International Program of Project 985, Sun Yat-Sen University, Raising Program of Major Project of Sun Yat-sen University (No. 1309089), and MOE Project of Key Research Institute of Humanities and Social Sciences at Universities, China (No. 13JJD720017).

References

1. L. Amgoud and H. Prade. Using arguments for making and explaining decisions. *Art. Int.*, 173(3-4):413–436, 2009.
2. E. Armengol and E. Plaza. Symbolic explanation of similarities in case-based reasoning. *Comput. Inf.*, 25(2-3):153–171, 2012.
3. K. D. Ashley. *Modeling legal arguments: Reasoning with cases and hypotheticals*. MIT press, 1991.
4. K. D. Ashley, C. Lynch, N. Pinkwart, and V. Aleven. A process model of legal argument with hypotheticals. In *Proc. JURIX*, pages 1–10, 2008.
5. K. Atkinson, T. Bench-Capon, and P. McBurney. Justifying practical reasoning. In *Proc. CMNA*, pages 87–90, 2004.
6. P. M. Dung, R. A. Kowalski, and F. Toni. Assumption-based argumentation. In *Argumentation in Artificial Intelligence*, pages 199–218. Springer, 2009.
7. P.M. Dung, R.A. Kowalski, and F. Toni. Dialectic proof procedures for assumption-based, admissible argumentation. *Art. Int.*, 170(2):114–159, 2006.
8. X. Fan, R. Craven, R. Singer, F. Toni, and M. Williams. Assumption-based argumentation for decision-making with preferences: A medical case study. In *Proc. CLIMA*, pages 374–390, 2013.
9. X. Fan and F. Toni. Decision making with assumption-based argumentation. In *Proc. TAFA*, pages 127–142, 2013.
10. J. Fox, D. Glasspool, V. Patkar, M. Austin, L. Black, M. South, D. Robertson, and C. Vincent. Delivering clinical decision support services: There is nothing as practical as a good theory. *J. of Biom. Inf.*, 43(5):831–843, 2010.
11. S. Heras, J. Jordn, V. Botti, and V. Julian. Argue to agree: A case-based argumentation approach. *Int. J. App. Reas.*, 54(1):82–108, 2013.
12. A Kakas and P Moraitis. Argumentation based decision making for autonomous agents. In *Proc. AAMAS*, pages 883–890, 2003.
13. A. T. Kronman. Precedent and tradition. *Yale Law J.*, 99(5):1029–1068, 1990.
14. C. Labreuche. A general framework for explaining the results of a multi-attribute preference model. *Art. Int.*, 175(7):1410–1448, 2011.
15. P. A. Matt, F. Toni, and J. Vaccari. Dominant decisions by argumentation agents. In *Proc. ArgMAS*, pages 42–59. Springer, 2009.
16. D. McSherry. Explanation in recommender systems. *Art. Int. Rev.*, 24(2):179–197, 2005.
17. J. Müller and A. Hunter. An argumentation-based approach for decision making. In *Proc. ICTAI*, pages 564–571, 2012.

18. F. S. Nawwab, T. J. M. Bench-Capon, and P. E. Dunne. A methodology for action-selection using value-based argumentation. In *Proc. COMMA*, pages 264–275, 2008.
19. F. Portet, E. Reiter, , A. Gatt, J. Hunter, S. Sripada, Y. Freer, and C. Sykes. Automatic generation of textual summaries from neonatal intensive care data. *Art. Int.*, 173(7):789–816, 2009.
20. H. Prakken. An abstract framework for argumentation with structured arguments. *Arg. and Comp.*, 1(2):93–124, 2010.
21. P. R. Quinlan, A. Thompson, and C. Reed. An analysis and hypothesis generation platform for heterogeneous cancer databases. In *Proc. COMMA*, pages 59–70, 2012.
22. E. Reiter, S. Sripada, J.Hunter, and I. Davy. Choosing words in computer-generated weather forecasts. *Art. Int.*, 167(1-2):137–169, 2005.