Available online at www.sciencedirect.com
SCIENCE DIRECT®
Artificial Intelligence 170 (2006) 925–952
Artificial Intelligence
www.elsevier.com/locate/artint

Generating and evaluating evaluative arguments

Giuseppe Carenini [a,*], Johanna D. Moore [b]

[a] *Computer Science Department, University of British Columbia, 2366 Main Mall, Vancouver, BC Canada V6T 1Z4*
[b] *Human Communication Research Centre, University of Edinburgh, 2 Buccleuch Place, Edinburgh, United Kingdom EH8 9LW*

Received 21 June 2005; received in revised form 15 May 2006; accepted 23 May 2006

Abstract

Evaluative arguments are pervasive in natural human communication. In countless situations people attempt to advise or persuade their interlocutors that something is desirable (vs. undesirable) or right (vs. wrong). With the proliferation of on-line systems serving as personal advisors and assistants, there is a pressing need to develop general and testable computational models for generating and presenting evaluative arguments. Previous research on generating evaluative arguments has been characterized by two major limitations. First, researchers have tended to focus only on specific aspects of the generation process. Second, the proposed approaches were not empirically tested. The research presented in this paper addresses both limitations. We have designed and implemented a complete computational model for generating evaluative arguments. For content selection and organization, we devised an argumentation strategy based on guidelines from argumentation theory. For expressing the content in natural language, we extended and integrated previous work in computational linguistics on generating evaluative arguments. The key knowledge source for both tasks is a quantitative model of user preferences. To empirically test critical aspects of our generation model, we have devised and implemented an evaluation framework in which the effectiveness of evaluative arguments can be measured with real users. Within the framework, we have performed an experiment to test two basic hypotheses on which the design of the computational model is based; namely, that our proposal for tailoring an evaluative argument to the addressee's preferences increases its effectiveness, and that differences in conciseness significantly influence argument effectiveness. The second hypothesis was confirmed in the experiment. In contrast, the first hypothesis was only marginally confirmed. However, independent testing by other researchers has recently provided further support for this hypothesis.

Keywords: Natural language generation; User tailoring; Preferences; Empirical evaluation

1. Introduction

Evaluative arguments are pervasive in natural human communication. In countless situations, people attempt to advise or persuade their interlocutors that something is desirable (vs. undesirable) or right (vs. wrong). For instance, doctors need to advise patients about which treatment is best for them. A teacher may need to convince a student that a certain course is (is not) the best choice for the student. And salespeople often need to compare similar products, explaining why one of the products would be more to the current customer's liking than the other(s). With the

[*] Corresponding author.
E-mail address: giuseppe.carenini@gmail.com (G. Carenini).

0004-3702/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.artint.2006.05.003

explosion of information available on-line and the ever-increasing availability of wireless devices, we are witnessing a proliferation of computer systems serving as personal assistants or advisors, e.g., [9,62], which aim to support or replace humans in similar communicative settings. The success of such systems will crucially depend on their ability to generate and present effective evaluative arguments.

In the 1990s, considerable research was devoted to developing computational models for automatically generating and presenting evaluative arguments. Several studies have investigated the process of selecting and structuring the content of an argument (e.g., [7,31,35,47]), and [23] developed a detailed model of how the selected content should be realized in natural language. Despite the abundance of prior work on this topic, the previous research has been characterized by two major limitations. First, because of the complexity of generating natural language, researchers have tended to focus only on specific aspects of the generation process. Second, because of a lack of systematic evaluation, it is difficult to gauge the effectiveness, scalability and robustness of the proposed approaches.

The research presented in this paper addresses these limitations. By following principles from argumentation theory and computational linguistics, we have developed a complete computational model for generating evaluative arguments. In our model, all aspects of the generation process are covered in a principled way, from selecting and organizing the content of the argument, to expressing the selected content in natural language. For content selection and organization, we devised an argumentation strategy based on guidelines from argumentation theory. For expressing the content in natural language, we extended and integrated previous work on generating evaluative arguments. The key knowledge source for both tasks is a quantitative model of user preferences. To empirically test critical aspects of our generation model, we have devised and implemented an evaluation framework in which the effectiveness of evaluative arguments can be measured with real users. The design of the evaluation framework was based on principles and techniques from several research fields, including computational linguistics, social psychology, decision theory and human computer interaction. Within the framework, we have performed an experiment to test two basic hypotheses on which the design of the computational model is based; namely, that tailoring an evaluative argument to a model of the addressee's preferences increases its effectiveness, and that differences in conciseness significantly influence argument effectiveness. The first hypothesis was only marginally confirmed in the experiment ($0.05 < p < 0.10$), while the second one was confirmed at $p < 0.05$. Moreover, recent work [62], which is a direct extension of our research, provided further independent empirical support for the first hypothesis.

In the next section, we focus on the problem of generating evaluative arguments tailored to a model of the user's preferences and we describe the design and development of our Generator of Evaluative Argument (GEA). In Section 2, we describe our evaluation framework. First, we justify the design of the evaluation framework by reviewing literature on persuasion from social psychology as well as previous work on evaluating natural language generation techniques. Next, we introduce and motivate the user task at the core of the framework. In particular, we illustrate how, in the context of this task, the effectiveness of an argument can be assessed by measuring its effects on user's behaviors, beliefs and attitudes. Section 3 describes the experiment we ran within the evaluation framework and in Section 4 we discuss related work on generating and evaluating evaluative arguments.

2. Generating evaluative arguments

The generation of evaluative arguments has been extensively investigated in the past. Yet, the computational models developed in previous work only cover sub-parts of the generation process. For instance, [35] provided a sophisticated approach only to content selection, while [23] was mainly limited to content realization. Furthermore, all earlier models were not informed by argumentation theory [42], a theory, rooted in rhetoric, providing guidelines on how effective arguments are to be generated.

In this section we present GEA, the first computational model that covers all aspects of generating evaluative arguments in a principled way, by effectively integrating general principles and techniques from argumentation theory and computational linguistics. GEA is a rather complex computational model. In this section, we describe its design and development in a top-down fashion. First, we illustrate how GEA specializes the pipeline architecture typically adopted in Natural Language Generation (NLG) systems and introduce the basic algorithms and knowledge structures. Then, we discuss a set of guidelines from argumentation theory on which an effective argumentation strategy can be based. After that, we introduce the quantitative model used in GEA to represent the user's preferences and describe an argumentation strategy that tailors the content as well as structure of an evaluative argument to such a model. The

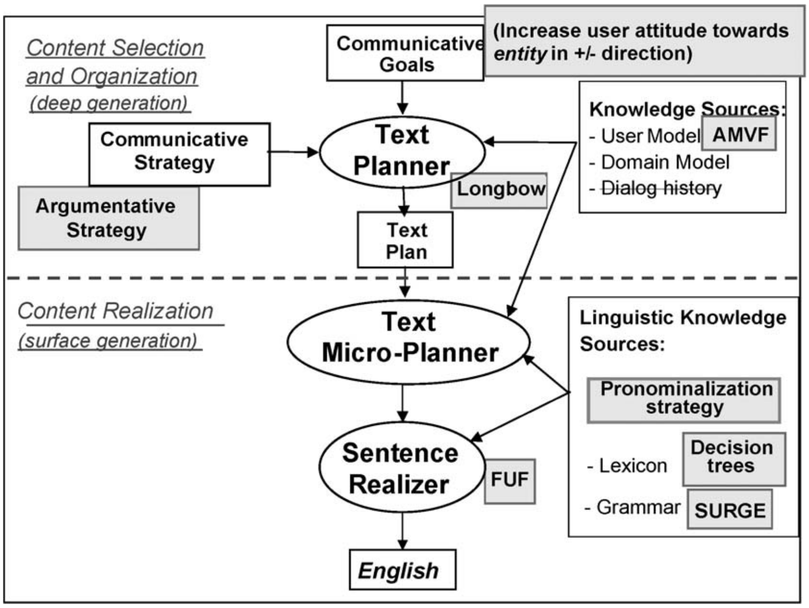


Fig. 1. The GEA architecture as a specialization of the generic NLG pipeline architecture.

section concludes with a detailed description of how GEA realizes the content selected by the argumentation strategy in natural language.

2.1. The architecture of the Generator of Evaluative Arguments (GEA)

Text generation involves two fundamental tasks: a process that selects and organizes the content of the text (deep generation), and a process that expresses the selected content in natural language (surface generation). GEA, like most previous work in NLG, makes the assumption that deep generation should strictly precede surface generation, and adopts the resulting pipeline architecture [50]. In this architecture (see center of Fig. 1 from top to bottom) a text planner selects and organizes content from a domain model by applying a communicative strategy to achieve a set of communicative goals, which are given as input. The output of text planning is a text plan, a data structure that specifies: the rhetorical structure of the text, the propositions that the text should convey and a partial order among those propositions. Then, a text Micro-Planner packages the selected content into sentences and selects words and syntactic structures to effectively express that content. Finally, a Sentence Realizer runs the output of the Micro-Planner through a computational grammar of English that produces English text. Notice that during both text planing and microplanning the content, the structure and the phrasing of the text can be tailored to a model of the communicative context (e.g., a user model).

Fig. 1 shows how GEA specializes the standard pipeline architecture for a generic NLG system. GEA specific features are shown as grey boxes in the figure and are in italics in the following text. The input to GEA is an abstract *evaluative communicative goal* expressing that the user attitude toward an entity in the domain of interest (e.g., a house in the real-estate domain) should be increased either in a positive or in a negative direction, with positive/negative meaning that the user should like/dislike the entity. Given an abstract communicative goal, the *Longbow* text planner [65] selects and arranges the content of the argument by applying a set of communicative strategies that implement an *argumentation strategy* based on guidelines for content selection and organization from argumentation theory (e.g., [42]). Two knowledge sources are involved in this process of goal and action decomposition (see Fig. 1): (i) A domain model representing entities and their relationships in a specific domain. (ii) An additive multi-attribute value function (*AMVF*), which is a decision-theoretic model of the user's preferences [14].[1]

[1] Currently, GEA is not a component of a dialogue system, so it is not sensitive to a dialogue history.

Next, the text plan is passed to the GEA microplanner which performs aggregation and lexicalization, and generates referring expressions. Aggregation, the packaging of semantic information into sentences, is performed according to standard techniques [50]. For lexicalization, the selection of lexical items to express the desired meaning, the GEA microplanner selects words to express evaluations by applying a *decision tree* that extends previous work on realizing evaluative statements [22]. Decisions about cue phrases (to express discourse relationships among text segments) are implemented as another *decision tree* based on features suggested in the literature (e.g., [36]). The generation of referring expressions in GEA is straightforward; an entity is always referred to either by its proper noun or by a pronoun. For pronominalization (deciding whether to use a pronoun or not to refer to an entity), a simple strategy based on *centering theory* [27] is applied. Finally, the output of text microplanning is unified by the GEA sentence realizer (FUF) with the Systemic Unification Realization Grammar of English (SURGE) [24].

We will now describe in detail the three key challenges in developing GEA: the design of the argumentation strategy, the development of the model of the users' preferences, and the design of the microplanner.

2.2. An argumentation strategy based on user preferences

2.2.1. Guidelines from argumentation theory

An argumentation strategy specifies what content should be included in the argument and how it should be arranged. This comprises several decisions: what represents supporting (or opposing) evidence for the main claim, where to position the main claim of the argument, what supporting (or opposing) evidence to include and how to order it, and how to order supporting and opposing evidence with respect to one another. Argumentation theory has developed guidelines specifying how these decisions can be effectively made (see [16,42,44,45] for details; see also [41] for an alternative discussion of some of the same guidelines). In this section, we describe the guidelines in detail. In Section 2.2.3 we will provide computational versions of these guidelines.

(a) *What represents supporting* (*or opposing*) *evidence for a claim and how to determine its strength*: Guidelines for this decision vary depending on the argument type. Limiting our analysis to evaluative arguments, argumentation theory indicates that supporting (or opposing) evidence and its strength should be determined according to a model of the reader's values and preferences. For instance, the risk involved in a game can be used as strong evidence for the claim that the reader should like the game, only if the reader likes risky situations a lot.

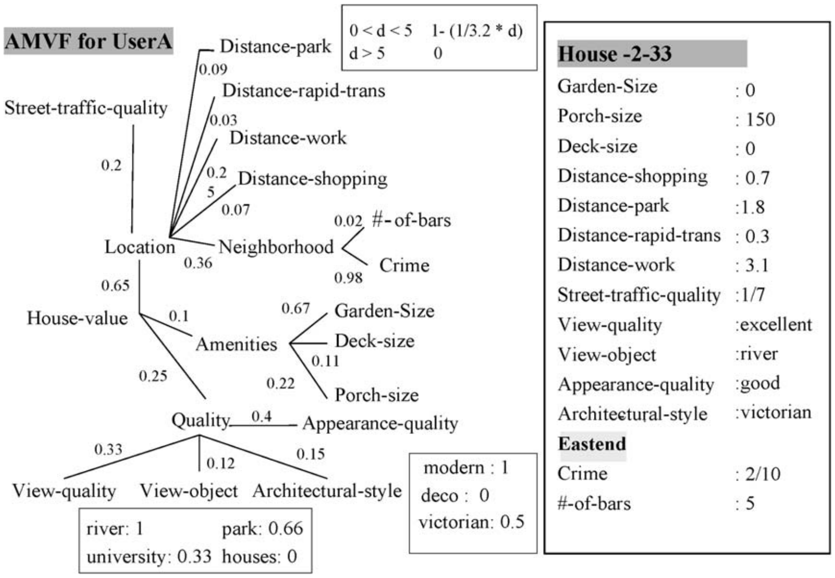
(b) *Positioning the main claim*: Claims are often presented up front, usually for the sake of clarity. Placing the claim early helps readers follow the line of reasoning. However, delaying the claim until the end of the argument can be effective, particularly when readers are likely to find the claim objectionable or emotionally shattering.

(c) *Selecting supporting* (*and opposing*) *evidence*: Often an argument cannot mention all of the available evidence, usually for the sake of brevity. Only strong evidence should be presented in detail, whereas weak evidence should be either briefly mentioned or omitted entirely.

(d) *Arranging/ordering supporting evidence*: Typically the strongest support should be presented first, in order to get at least provisional agreement from the reader early on. If at all possible, at least one very effective piece of supporting evidence should be saved for the end of the argument, in order to leave the reader with a final impression of the argument's strength. This guideline, proposed in [42], is a compromise between the climax and the anti-climax approaches discussed in [43].

(e) *Addressing and ordering the counterarguments* (*opposing evidence*): There are three options for this decision: not to mention any counterarguments, to acknowledge them without directly refuting them, to acknowledge them and directly refuting them. Weak counterarguments may be omitted. Stronger counterarguments should be briefly acknowledged, because that shows awareness of the issue's complexity, as well as a reasonable and broad-minded attitude. A counterargument, once acknowledged, may also need to be refuted, if the reader agrees with a substantially different position. Finally, counterarguments should be ordered to minimize their effectiveness: strong ones should be placed in the middle, weak ones upfront and at the end.

(f) *Ordering supporting and opposing evidence*: A preferred ordering between supporting and opposing evidence appears to depend on whether the reader is aware of the opposing evidence. If so, the preferred ordering is opposing before supporting, and the reverse otherwise.



House -2-33	
Garden-Size	: 0
Porch-size	: 150
Deck-size	: 0
Distance-shopping	: 0.7
Distance-park	:1.8
Distance-rapid-trans	: 0.3
Distance-work	: 3.1
Street-traffic-quality	:1/7
View-quality	:excellent
View-object	:river
Appearance-quality	:good
Architectural-style	:victorian
Eastend	
Crime	: 2/10
#-of-bars	: 5

Fig. 2. Sample AMVF preference model for UserA (left). Information about sample House-2-33 (right).

Although these guidelines provide information about the types of content to include in an evaluative argument and how to arrange this content, the design of a computational argumentative strategy requires that the concepts mentioned in the guidelines be formalized in a coherent computational framework. In particular, we require a model of the reader's values and preferences that will allow a system to: identify supporting and opposing evidence (guideline a); operationally define the term "objectionable claim" (guideline b) using a measure of the discrepancy between the reader's initial position and the argument's main claim;[2] measure the strength of supporting or opposing evidence (guidelines c, d, and e); and represent whether the reader is aware of certain facts (guideline f). We describe such a model in the next section.

2.2.2. *Modeling user preferences with MAUT: AMVFs*

One model that satisfies the requirements noted above is the additive multiattribute value function (AMVF), which is based on multiattribute utility theory (MAUT) [14]. MAUT is widely used in decision theory, where it was originally developed, and has become a common choice in the field of artificial intelligence and intelligent user interfaces [2,31,39]. Models similar to AMVF have also proven useful in psychology, in particular for the study of consumer behavior [56]. A critical aspect of AMVFs is that they can be elicited from people in a reliable and efficient way [1,21].

As their name suggests, multi-attribute utility models are based on the notion that if something is valued, it is valued for multiple reasons [34]. An AMVF is a model of an individual's values and preferences with respect to entities in a given class. To build an AMVF for a particular domain, we must identify the attributes that contribute to users' overall assessment of entities, and determine the relative importance of each attribute for particular users.

More formally, an AMVF consists of a *value tree* and a set of *component value functions*. A value tree is a decomposition of an entity's value into a hierarchy of aspects of the entity, called *objectives* in decision theory, in which the leaves of the value tree correspond to *primitive objectives*. For example, Fig. 2 (left side) shows a value tree for the real estate domain, in which the value of a house is a combination of the values of its *location*, *amenities*, and *quality*. The value for amenities is further broken down into values for the primitive objectives *garden-size*, *porch-size* and *deck-size*. *Location* and *quality* are also broken down into more primitive objectives. A component value function for a primitive objective expresses the preferability of each value for that objective as a number in the interval [0, 1], with the most preferable value mapped to 1, and the least preferable value to 0.[3] For instance, in Fig. 2 the value *modern*

[2] An operational definition for "emotionally shattering" is outside the scope of this work.

[3] For illustration, three component value functions are shown (as text boxes) in Fig. 2.

of the primitive objective *architectural-style* is the most preferred by UserA, and a *distance-from-park* of 1 mile has preferability $(1-(1/3.2*1))=0.69$.

The arcs in the value tree are weighted to represent how valuable it would be for the decision maker to move from the worst to the best level of an objective (with respect to doing the same for its siblings). For instance in Fig. 2, UserA would consider moving from a *deco* house to a *modern* house as slightly more valuable than moving from a house with a *view on other houses* to a house with a *view on a river* (cf. weight for *Architectural-style* is 0.15, whereas weight for *View-object* is 0.12).

Formally, an AMVF predicts the value $v(e)$ of an entity e as follows:

$$v(e)=v(x_1,\dots,x_n)=\sum_{i=1}^{n} w_i v_i(x_i),$$

where

- $(x_1,\dots,x_n)$ is the vector of primitive objective values for an entity e,
- $\forall$ primitive objective i, v_i is the component value function and w_i is its weight, with $0\leqslant w_i\leqslant 1$ and $\sum_{i=1}^{n} w_i=1$; w_i is equal to the product of all the weights on the path from the root of the value tree to the primitive objective i.

A function $v_o(e)$ can also be defined for each objective. When applied to an entity, this function returns the value of the entity with respect to that objective. For instance, assuming the value tree shown in Fig. 2, we have:

$$\begin{aligned} v_{Quality}(e) = {} & \big(w_{View\text{-}Quality} * v_{View\text{-}Quality}(e)\big) \\ & + \big(w_{View\text{-}Object} * v_{View\text{-}Object}(e)\big) \\ & + \big(w_{Architectural\text{-}Style} * v_{Architectural\text{-}Style}(e)\big) \\ & + \big(w_{Appearance\text{-}Quality} * v_{Appearance\text{-}Quality}(e)\big). \end{aligned}$$

Thus, given an AMVF for a particular user, it is possible to compute how valuable an entity is to that individual. Furthermore, it is possible to compute how valuable any objective (i.e., any aspect of that entity) is for that person. All of these values are expressed as a number in the interval [0, 1].

In general, when uncertainty is present, a user's valuation of an entity can be represented as a linear combination of her preferences for the primitive objectives (i.e., as an AMVF) only in cases where these preferences satisfy the condition of additive independence. That is, each objective is assumed to be independent of all the others. However, standard heuristic tests with users have shown that additive models are a good approximation of people's preferences under conditions of certainty [21]. Thus, AMVFs can safely be used either in situations with no uncertainty or in uncertain situations once additive independence has been verified.

Edwards and Barron [21] have shown that AMVFs can be elicited from people in a reliable and efficient way. They devised SMARTER, a simple procedure for eliciting objective weights and component value functions from users.

Objective weights are user-specific, reflecting individual preferences about tradeoffs between entities in the domain. To elicit the weights, SMARTER asks the user to perform a series of easy assessments, $N-1$ for N objectives, from which a user-specific ranking of the objectives can be generated. From such a ranking, the weights can be computed according to the following formula, which specifies the weight of the kth objective as:

$$w_k=(1/K)\sum_{i=k}^{K}(1/i).$$

There is considerable experimental evidence indicating that simple attribute ranking is both more efficient than, and nearly as accurate as, traditional more time-consuming methods in which weights are directly elicited from users. Traditional methods require $k*(N-1)$ assessments for N objectives, with k possibly quite large [14]. Moreover, the resulting efficiency gain does not appear to penalize accuracy. Simulation studies have shown that SMARTER introduces only a 2% utility loss [1].

The elicitation of component value functions in SMARTER is simplified as follows. If the function specifies the preferability of a continuous objective (e.g., *garden-size*) the user needs only to choose among essentially three basic possibilities: the function either increases linearly across the whole value range, decreases linearly across the whole value range, or increases linearly on a sub-interval and then decreases linearly on its complement. If the function specifies the preferability of a discrete objective (e.g., *architectural-style*), it can be acquired in a manner similar to the way in which weights are elicited, i.e., by having the user rank all the possible values.

SMARTER makes several simplifying assumptions. Nevertheless, it is remarkably effective. Ref. [12] shows that models developed using SMARTER return evaluations that correlate highly with experts' holistic judgements (Pearson's coefficient 0.68; $p < 0.001$).

2.2.3. Operationalizing the argumentation strategy

Presenting an evaluative argument is an attempt to persuade the reader that a value judgement applies to an entity. The value judgement, also called the argumentative intent, can either be positive (in favor of the subject), or negative (against the subject).[4] The subject can be a single entity (e.g., "Ulysses is a very good book"), the difference between two entities (e.g., "Vancouver is somewhat better than Seattle"), or any other form of comparison among entities in a set (e.g., "Vancouver is the best city in North America"). We now describe how we can use the information in an AMVF to operationalize the guidelines presented in Section 2.2.1.

Guideline (a): Given a user's AMVF, it is straightforward to establish what represents supporting or opposing evidence for an argument with a given argumentative intent and a given subject. If the argumentative intent is positive, objectives for which the subject has positive value can be used as supporting evidence, whereas objectives for which the subject has a negative value can be used as opposing evidence (the opposite holds when the argumentative intent is negative). The value of different subjects can be reasonably measured as follows. If the subject is a single entity e, the value of the subject for an objective o is $v_o(e)$, and it is positive when it is greater than 0.5, the midpoint of [0, 1] (negative otherwise). In contrast, if the subject is a comparison between two entities (e.g., $v(e_1) > v(e_2)$), the value of the subject for an objective o is $[v_o(e_1) - v_o(e_2)]$, and it is positive when it is greater than 0 (negative otherwise).

Guideline (b): Since argumentative intent is a value judgement, we can reasonably assume that instead of being simply positive or negative, it may be specified more precisely as a number in the interval [0, 1] (or as a specification that can be normalized to a value in this interval). Then, the term "objectionable claim" can be operationally defined. If we introduce a *measure of discrepancy* (*MD*) as the absolute value of the difference between the argumentative intent and the reader's expected value of the subject before the argument is presented (based on her AMVF), a claim becomes more and more "objectionable" for a reader as *MD* moves from 0 to 1.

Guidelines (c) (d) (e): The strength of the evidence in support of (or opposition to) the main argument claim is critical in selecting and organizing the argument content. To define a measure of the strength of support (or opposition), we adopt and extend previous work on explaining decision theoretic advice based on an AMVF. Klein [35] presents explanation strategies (not based on argumentation theory) to justify the preference of one alternative from a pair. In these strategies, the *compellingness* of an objective measures the objective's strength in determining the overall value difference between the two alternatives, other things being equal. And an objective is *notably-compelling?* (i.e., worth mentioning) if it is an outlier in a population of objectives with respect to compellingness. The formal definitions are:

$$\mathit{compellingness}(o, a_1, a_2) = w(o, \mathit{root})\big|\big[v_o(a_1) - v_o(a_2)\big]\big|,$$

where

- o is an objective, a_1 and a_2 are alternatives,
- $w(o, \mathit{root})$ is the product of the weights of all the links on the path from o to the *root* objective of the value tree,
- v_o is the component value function for leaf objectives (i.e., attributes), and it is the recursive evaluation over *children*(o) for non-leaf objectives ($w(o, o')$ is the weight on the link from o' to o):

$$v_o(a) = \sum_{o' \in \mathit{children}(o)} w(o, o')v_{o'}(a),$$

$$\mathit{notably\text{-}compelling?}(o, \mathit{opop}, a_1, a_2) \equiv \mathit{compellingness}(o, a_1, a_2) > \mu_x + k\delta_x,$$

[4] Arguments can also be neutral. However, in this paper we do not discuss arguments with a neutral argumentative intent.

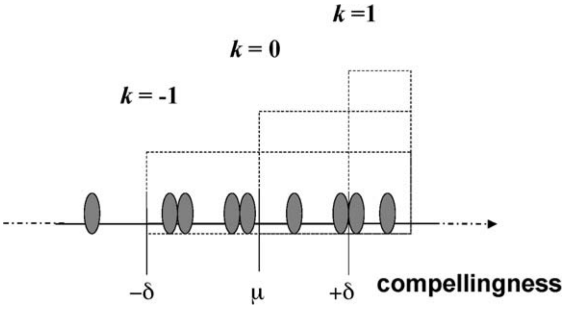


Fig. 3. Sample population of objectives represented by dots and ordered by their compellingness.

where

- o, a_1 and a_2 are defined as in the previous definition; *opop* is an objective population (e.g., *siblings*(o)), and $|opop| > 2$,
- $X = \{x = compellingness(p, a_1, a_2) \mid p \in opop\}$,
- μ_x is the mean of X, δ_x is the standard deviation.

We have adopted *compellingness* as our measure of the strength for supporting or opposing evidence. We have adopted *notably-compelling?* as a decision criterion for including a piece of evidence in the argument. Notice that the definition of *notably-compelling?* relies on a constant k which determines a lower bound of *compellingness* for an objective to be included in an argument. So, by setting the constant k to different values, it is possible to control, in a principled manner, the number of objectives (i.e., pieces of evidence) that are included in an argument, thus controlling the degree of conciseness of the generated arguments. As shown in Fig. 3, for $k = 0$ only objectives with *compellingness* greater than the average *compellingness* in a population are included in the argument (4 in the sample population). For higher positive values of k fewer objectives are included (only 2, when $k = 1$), and the opposite happens for negative values (8 objectives are included, when $k = -1$).

The concepts *compellingness* and *notably-compelling?* were defined to support arguments that one entity is more valuable than another. We have defined similar measures for arguing the value of a single entity, which we have termed *s-compellingness* and *s-notably-compelling?*. An objective can be *s-compelling* either because of its strength or because of its weakness in contributing to the value of an alternative. So, if m_1 measures how much the value of an objective contributes to the overall value difference of an alternative from the worst possible case[5] and m_2 measures how much the value of an objective contributes to the overall value difference of the alternative from the best possible case, we define *s-compellingness* as the greatest of the two quantities m_1 and m_2. Following the terminology introduced in the two previous equations we have:

$$s\text{-}compellingness(o, a) = w(o, root) * max\big[v_o(a), \big[1 - v_o(a)\big]\big].$$

We give to *s-notably-compelling?* a definition analogous to the one for *notably-compelling?*

$$s\text{-}notably\text{-}compelling?(o, opop, a) \equiv s\text{-}compellingness(o, a) > \mu_x + k\delta_x.$$

In *s-notably-compelling?* the constant k plays the same role as in *notably-compelling?*: by setting the constant k to different values, it is possible to control the degree of conciseness of the generated arguments.

Guideline (f): An AMVF does not represent whether the reader is or is not aware of certain facts. We assume this information is represented elsewhere.

2.2.4. The argumentation strategy

We have applied the guidelines from argumentation theory and the corresponding formal definitions described in the previous section to develop the argumentative strategy shown in Fig. 4. The steps in the strategy are marked with

[5] a_{worst} is an alternative such that $\forall o\ v_o(a_{worst}) = 0$, whereas a_{best} is an alternative such that $\forall o\ v_o(a_{best}) = 1$.

```
Argue(subject, Root, ArgInt, $k$)
;; (A) assignments
If subject = single-entity = $e$ then $SV_{o_i} = v_{o_i}(e)$
            Measure-of-strength = s-compellingness
            Worth-mention? = s-notably-compelling?
Else If subject = $e_1, e_2$ then $SV_{o_i} = [v_{o_i}(e_1) - v_{o_i}(e_2)]$
            Measure-of-strength = compellingness
            Worth-mention? = notably-compelling?
;; (B) content selection
  Eliminate all objectives $o_i | \neg$Worth-mention? ($o_i$, $siblings(o_i)$, $subject$, $Root$)      ;guideline(c)
  $AllEvidence \leftarrow children(Root)$
  $AllInFavor \leftarrow$ all $o | o \in AllEvidence \wedge (SV_o \approx ArgInt)$      ;guideline(a)
  $SecondBestObjInFavor \leftarrow$ second most compelling objective $o | o \in AllInFavor$
  $RemainingObjectivesInFavor \leftarrow AllInFavor - SecondBestObjInFavor$
  $ContrastingObjectives \leftarrow AllEvidence - AllInFavor$      ;guideline(a)

;; (C) ordering constraints
  AddOrdering($Root \prec AllEvidence$) ;; we assume MD = 0, so claim is not objectionable      ;guideline(b)
  If Aware(User, $ContrastingObjectives$) then      ;guideline(f)
            AddOrdering($ContrastingObjectives \prec AllInFavor$)
  Else AddOrdering($ContrastingObjectives \succ AllInFavor$);
  AddOrdering($RemainingObjectivesInFavor \prec SecondBestObjInFavor$)      ;guideline(d)
  Sort($RemainingObjectivesInFavor$; decreasing order according to Measure-of-strength)      ;guideline(d)
  Sort($ContrastingObjectives$; strong ones in the middle, weak ones upfront and at the end)      ;guideline(e)

;; (D) steps for expressing or further argue the content
  Express-Value(subject, Root, ArgInt)
  For all $o \in ContrastingObjectives$, Express-Value(subject, $o$, $SV_o$)      ;guideline(e)
  For all $o \in AllInFavor$, If $\neg$ leaf($o$) then Argue(subject, $o$, $SV_o$, $k$)
                        Else Express-Value(subject, $o$, $SV_0$)

Legend: $(a \prec b) \leftrightarrow$ a preceeds b
        $(v_1 \approx v_2) \leftrightarrow v_1$ and $v_2$ are both positive or negative values
            (see Section guideline (a) for what this means for different subjects)
        $SV_{o_i}$: function computing the subject value for objective $o_i$
```

Fig. 4. The argumentation strategy.

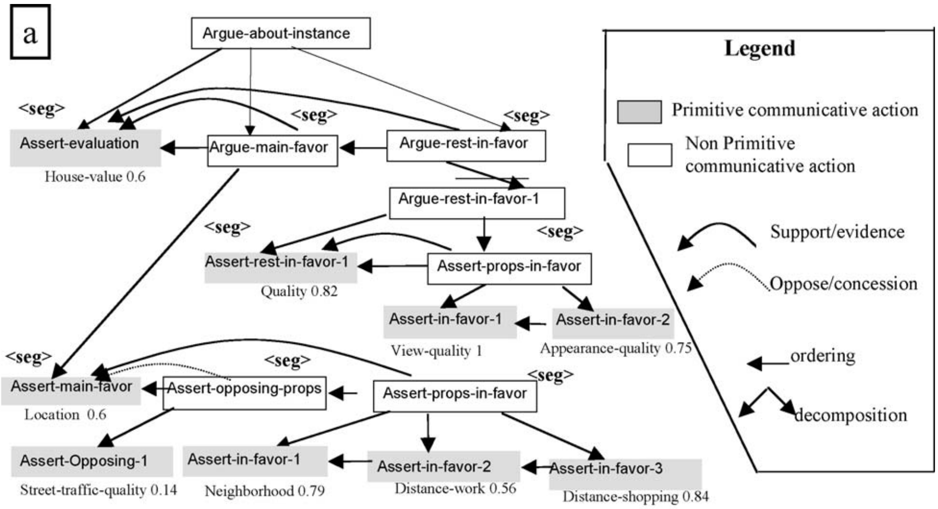
the guideline they are based on. The strategy is designed for generating arguments which present an evaluation of the subject equivalent to the one that the reader would be expected to hold given her model of preferences (i.e., the argumentative intent is equal to the expected value, so $MD = 0$).[6]

We now examine the strategy in detail, after introducing necessary terminology. The strategy is called *Argue* and takes four arguments (first line in Fig. 4). The *subject* is the entity (or entities) that is to be evaluated or compared, and can be either a single entity or a pair of entities in the domain of interest. *Root* can be any objective in the value tree for the evaluation (e.g., the overall value of a house, its location, its amenities). *ArgInt* is the argumentative intent of the argument, a number in [0, 1], with 0 meaning the worst and 1 the best. The constant k, part of the definitions of *notably-compelling?* and *s-notably-compelling?*, determines the degree of conciseness of the argument.

The *Express-Value* function, used at the end of the strategy, indicates that the objective applied to the subject must be realized in natural language with a certain argumentative intent.

In the first part of the strategy (A in the figure), depending on the nature of the subject, an appropriate measure of evidence strength is assigned, along with the appropriate predicate that determines whether a piece of evidence is worth mentioning. After that (in part (B)), only evidence that is worth mentioning is assigned as supporting or opposing evidence by comparing its value to the argument intent. In part (C) ordering constraints from argumentation

[6] An alternative strategy, for generating arguments whose argumentative intent was greater (or lower) than the expected value could also be defined in our framework. This strategy would boost the evaluation of supporting evidence and include only weak counterarguments, or omit them entirely (the opposite if the target value was lower than the expected value).



b

<< House 2-33 is an interesting house. > b_1<< In fact, it is in a reasonable location in the safe Eastend neighborhood. > b_2<Even though the traffic is intense on 2nd Street, > b_3< House 2-33 is reasonably close to work. And it also offers easy access to the shops. >> b_4<<Furthermore, the quality of House 2-33 is good. > b_5< House 2-33 offers an excellent view. And it also looks beautiful. >>>

Fig. 5. Text plan, segmentation structure and corresponding argument generated by GEA about House-2-33, tailored to UserA with $k = -0.3$. UserA's model and the information about House-2-33 are shown in Fig. 2.

theory are applied. Notice that we assume a predicate *Aware* that is true when the user is aware of a certain fact, false otherwise.

Finally, in part (D), the argument claim is expressed in natural language by calling the ExpressValue function, which realizes the objective Root applied to subject with the argument intent ArgInt. Then, the opposing evidence (i.e., ContrastingSubObjectives), that must be considered, but not in detail, is also expressed in natural language. In contrast, supporting evidence is presented in detail, by recursively calling the strategy on each supporting piece of evidence whose corresponding objective is not a leaf of the value tree.

The argumentation strategy has been implemented as a library of communicative action decompositions for the Longbow discourse planner [66]. As shown in Fig. 1, the application of the argumentation strategy produces a text plan for an evaluative argument tailored to a given user. For instance, when the argumentation strategy is applied to the preference model and the entity introduced in Fig. 2, (i.e., subject = House-2-33, Root = HouseValue for UserA), with ArgIntent = 0.6, and $k = -0.3$, the text plan shown in Fig. 5(a) is generated.

The leaves of the text plan express the propositions that the argument should convey (e.g., ⟨*Assert that the Quality of House-2-33 has for UserA a value 0.82*⟩[7]). Note that only a subset of the objectives in the original AMVF are included in the plan. These are the *s-notably-compelling* objectives that were selected in part (B) of the strategy. The nodes of the text plan (i.e., the communicative actions) are also ordered (e.g., the action Assert-opposing-props should be performed before Assert-props-in-favor). These ordering relations were established in part (C) of the strategy. Finally, the text plan specifies the rhetorical structure of the argument. This structure is expressed by the hierarchical decomposition of the text plan and the rhetorical relations of *evidence* and *concession* between elements in the hierarchy. The plan hierarchical decomposition is generated in part (D) of the strategy by the recursive calls, while the rhetorical relations are determined in part (B) (guideline (a)).

[7] As in the AMVF, values are expressed in the [0,1] interval, with 1 corresponding to best possible, and the 0 to worst possible.

2.3. GEA microplanner

The output of the argumentation strategy is a text plan indicating the propositions to include in the argument and the overall structure the argument should take. This text plan is then passed to the microplanner (see Fig. 1) which performs aggregation, lexicalization and referring expression generation. To illustrate GEA's microplanner, we will describe how the sample text plan in Fig. 5(a) is realized into the corresponding argument shown in Fig. 5(b).[8]

A key aspect of the text plan for microplanning is the specification of discourse segments. Note that each node in the plan that participates in a rhetorical relation corresponds to a discourse segment. In Fig. 5(a), each discourse segment is marked with a ⟨*seg*⟩ label and corresponding portions of text in Fig. 5(b) are enclosed in angle brackets. Segments are internally structured and consist of an element that most directly expresses the discourse purpose of the segment (the element at the head of the relation arrows in the figure) and any number of constituents supporting that purpose.[9] We now illustrate each microplanning task in detail.

Lexicalization proper (i.e., no discourse cue selection): Lexicalization is the task of selecting words and associated syntactic structures to express semantic information. The GEA microplanner performs a simple form of lexical choice. For each proposition in the text plan it chooses the most appropriate proto-phrase to express that proposition. This decision is based on the objective of the proposition and its value for the current user. For example, in the sample text plan in Fig. 5(a), the proposition (*Location House-2-33 0.6*) is mapped to a proto-phrase which (after pronominalization) is realized as "it has a reasonable location", while the proposition (*Distance-shopping House-2-33 0.84*) is mapped to a proto-phrase which is realized as "it offers easy access to the shops". Mapping to proto-phrases is implemented by decision trees. Fig. 6(top) shows a portion of the decision tree for mapping the objectives to proto-phrases. Because there is no linguistic theory indicating how to realize the numeric intervals in natural language, we have based the choice of adjectives (e.g., "reasonable", "excellent") on our own estimates.

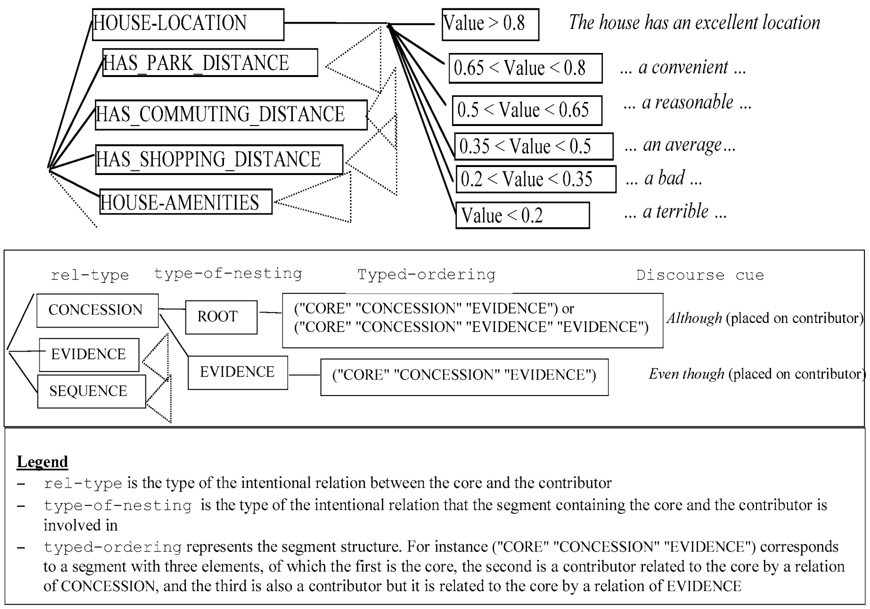
In practice, lexicalization proper in GEA is implemented in the Functional Unification Framework (FUF) by extending previous work on realizing evaluative statements [22]. The decision tree, partially shown in Fig. 6(top), is represented as a FUF grammar, while the selection and instantiation of the template to express a proposition is performed by unifying that proposition with the grammar.

Aggregation: Aggregation is the task of packaging semantic information into sentences. Three basic types of aggregation can be identified [50]. In *simple conjunction*, two or more informational elements are combined within a single sentence by using a connective such as "and". For instance, two informational elements that could be realized independently as (a_1) "House B-11 is far from a shopping area" and (a_2) "House B-11 is far from public transportation" can be combined and realized as the single sentence "a_1 and a_2". In *conjunction via shared participants*, two or more informational elements that share argument positions and are filled with the same content are combined to produce a surface form where the shared content is realized only once. For instance, the two informational elements aggregated above in a simple conjunction could be combined in a conjunction via shared participants as "House B-11 is far from a shopping area and public transportation". Finally, in *syntactic embedding*, an informational element that might have been realized as a separate major clause is instead realized as a constituent embedded into some other realized element. For instance, two informational elements that could be realized independently as "House B-11 offers a nice view" and "House B-11 offers a view of the river" can be combined and realized as "House B-11 offers a nice view of the river".

GEA performs both aggregation via shared participants and by syntactic embedding. To ensure argument coherence, aggregation is only attempted between objectives that are related to a claim by the same rhetorical relation type (evidence or concession). This is consistent with a heuristic proposed in [55] which addresses the question of when aggregation is (not) desirable, namely, "Do not express more than one type of rhetorical relation within a single sentence." In our example, we have only one aggregation between the *Location* and *Neighborhood* objectives, which are both related to the main claim by a relation of type evidence. Note that (*Neighborhood* is related to the main claim

[8] The text plan does not include the objective Crime (which is included in the argument). The reason is that the current implementation of the argumentation strategy only processes objectives of depth < 3 in the AMVF. The objective Crime is reintroduced in the argument by subsequent processing in an ad hoc fashion.

[9] The main element and the supporting elements are given different names in different discourse theories. In this paper we call the main component of a relation the core and the supporting elements the contributors.



Legend

- `rel-type` is the type of the intentional relation between the core and the contributor
- `type-of-nesting` is the type of the intentional relation that the segment containing the core and the contributor is involved in
- `typed-ordering` represents the segment structure. For instance ("CORE" "CONCESSION" "EVIDENCE") corresponds to a segment with three elements, of which the first is the core, the second is a contributor related to the core by a relation of CONCESSION, and the third is also a contributor but it is related to the core by a relation of EVIDENCE

Fig. 6. Portions of the decision trees used by the micro-planner for lexical choice (top) and for discourse cue selection (bottom).

through a chain of two evidence relations (see plan in Fig. 5(a)). The two propositions are aggregated by syntactic embedding.

Cue phrase generation: Although substantial research effort has been devoted to the development of a computational theory of discourse cue usage (e.g., [19,29,36,37,61]), a comprehensive theory is still lacking, and many open questions remain. Nevertheless, there is considerable consensus in the field about what factors may influence the usage of discourse cues. These include features that characterize the relationship (e.g., intentional, informational, syntactic) between the core and the contributor, features of the segment structure in which the core and contributor appear, and features related to the embedding within or outside a segment. Lacking a comprehensive theory, developers of NLG systems typically follow the methodology of devising, for the genre of interest, a specialized algorithm which relies on a carefully selected subset of these features [50]. In GEA, cue phrase selection and placement are implemented as a decision tree taking into account the following features, which have been suggested in the literature: (a) the intentional relationship between the core and the contributor, (b) the overall structure of the segment in which core and contributor appear (including the position of core and contributor(s) within the segment), and (c) the relationship in which the core and contributor segment itself is involved. For example, if by applying the portion of the decision tree shown in Fig. 6(bottom) to the text plan for our example, the reader can verify why "Even though" is used to mark the only concession in our sample argument.

Pronominalization: Most pronominalization algorithms in NLG rely on the notion of the "focus" or "center" of a sentence [50]. GEA decides whether to use a pronoun to refer to the evaluated entity by applying a simple strategy inspired by centering [27]. Centering theory indicates that the entity providing a link to the previous discourse in a locally coherent discourse (i.e., a discourse segment) should preferentially be realized as a pronoun rather than a repeated definite description.[10] Since in GEA the entity providing a link to the previous discourse is always the entity being evaluated by the argument, a straightforward application of centering would imply that within a discourse segment successive references to that entity are realized as pronouns, while at the beginning of a new segment a definite description is used to mark the segment boundary. For the arguments generated by GEA, we noticed that although the centering-based pronominalization policy works well for references within a segment, it is too restrictive for references at a segment boundary. In particular, it appears that a definite description at the beginning of a new segment is cumbersome and unnecessary when the segment boundary is already explicitly marked by a discourse

[10] This thesis has been empirically verified in [25].

cue, and a pronoun has not been used to refer to the entity in the previous sentence. So, GEA realizes the entity as a pronoun not only within a segment, but also in the situation just described. Clearly, the application of this strategy requires information about how the argument will be segmented in the final text. As described previously, the text plan expresses text segmentation: the core or contributor of each rhetorical relation corresponds to a segment. In Fig. 5(a), each discourse segment is marked with a $\langle seg \rangle$ label. The corresponding text in Fig. 5(b) contains five segment boundaries: $[b_1 \ldots b_5]$. For illustration, the pronominalization strategy is applied to the segment boundary b_1 as follows: b_1 is explicitly marked by the discourse cue "in fact" and a pronoun has not been used in the sentence preceding b_1 to refer to House-2-23. Thus, a pronoun is used to refer to that entity in the following sentence.

2.4. GEA: Summary and portability

GEA has been implemented as a complete and modular NLG system for generating user tailored evaluative arguments. We have seen how GEA covers all aspects of the process of generating evaluative arguments, from selecting and organizing the content of the argument, to expressing the selected content in natural language. For content selection and organization, GEA applies an argumentation strategy based on guidelines from argumentation theory. To express the content in natural language, GEA relies on a set of techniques that extend and integrate previous work in computational linguistics on microplanning and realizing evaluative arguments. Finally, a quantitative model of the user preferences expressed as an AMVF is the key knowledge source used by GEA in tailoring the content, organization and phrasing of the generated arguments to its users. The GEA implementation is largely domain independent. To port the system to a new domain, the implementor needs only to specify: (i) a value decomposition for each relevant entity (i.e., an AMVF in which the weights and the component value functions are not specified). (ii) a decision tree for lexicalization proper, like the one partially shown in Fig. 6(bottom), and (iii) an indication of which objectives can be aggregated and what type of aggregation is allowed (e.g., distance objectives can be aggregated by conjunction via shared participants).

3. Evaluating evaluative arguments

The goal of the work presented in this paper is to complete a research cycle that starts by designing a computational model and ends by empirically testing the design of the model. In the previous sections, we described GEA, a computational model for generating evaluative arguments tailored to the user's preferences. In this section we present an evaluation framework in which the effectiveness of the evaluative arguments generated by GEA can be measured with real users. Our framework is based on principles and techniques from social psychology, NLG, decision theory and human computer interaction. We first discuss literature from social psychology on persuasion and argument effectiveness. Next, we examine previous work on evaluating NLG techniques. Finally, we describe the architecture of our evaluation framework and its rationale.

3.1. Research in psychology on persuasion and argument effectiveness

In social psychology and communication theory, attitudes, beliefs and persuasion are defined as follows [45]:

- *Attitudes* are evaluative tendencies regarding some feature of the environment that are typically phrased in terms of like and dislike or favor and disfavor.
- *Beliefs* are assessments that something is or is not the case.
- *Persuasion* involves an intentional communicative act that attempts to affect the current or future behavior of the addressees by creating, changing or reinforcing addressees' attitudes and beliefs.

The focus of this work is on evaluative arguments that attempt to change or reinforce users' attitudes (vs. beliefs). Thus, beliefs will not be discussed further in this paper.

Since the goal of evaluative arguments is to affect behavior by affecting attitudes, it follows that their effectiveness can be tested by comparing measurements of subjects' attitudes or behavior before and after their exposure to the argument. For instance, to compare the effectiveness of two arguments that, by positively evaluating the state of being fit, attempt to change a person's amount of daily exercise (a behavior), we can perform an experiment in which we

(a) How would you judge house B-11?
(The more you like the house the closer you should put a cross to *"good choice"*.)
bad choice : ___ : ___ : ___ : ___ : ___ : ___ : ___ : ___ : ___ : *good choice*

(b) How sure are you that you have selected the best house for you among the ones available?
unsure : ___ : ___ : ___ : ___ : ___ : ___ : ___ : ___ : ___ : *sure*

Fig. 7. Sample self-reports.

compare the amount of daily exercise in two groups of subjects before and after participants have been exposed to one or the other of the evaluative arguments.

In many experimental situations, however, measuring effects on overt behavior can be problematic, and therefore research on persuasion is often based on measurements of attitudes or declarations of behavioral intentions [45]. The most common technique for measuring attitudes is the semantic differential self-report, in which subjects are presented with a scale whose endpoints are bipolar terms (e.g., "good choice" vs. "bad choice"), usually separated by seven or nine equal spaces that participants use to evaluate an attitude or belief statement (see Fig. 7 for examples).

Ref. [45] also suggests that some individuals may be naturally more resistant to persuasion than others, and thus individual differences must be taken into account when studying persuasion. Features of individuals that seem to be important in this respect are: argumentativeness (tendency to argue), intelligence, self-esteem, and the need for cognition (tendency to engage in and to enjoy effortful cognitive endeavours) [4,30]. Therefore, it is crucial to control for these variables when attempting to evaluate the persuasiveness of an argument.

Finally, an argument can also be evaluated by the addressee with respect to dimensions of quality, such as coherence, content, organization, writing style and convincingness. However, evaluations based on judgements along these dimensions are clearly weaker than evaluations measuring actual attitudinal and behavioral change [48].

3.2. Evaluation of NLG models

Three main methods have been proposed in the literature for the purpose of evaluating approaches to natural language generation: human judges, corpus-based evaluation, and task efficacy. All have their shortcomings, and it is important to choose the appropriate method to test the hypothesis that one is interested in. We now present a critical overview of the three methods and clarify why task efficacy is the most appropriate for testing the effectiveness of evaluative arguments that are tailored to a model of the user's preferences.

The *human judges* evaluation method requires a panel of judges to score outputs of a number of different generation models [5,6,17,40,59]. To compare models, each judge in a panel is given outputs generated by each of the models, and asked to rate the outputs on dimensions of text quality, such as coherence, content, organization, writing style and correctness. Note that the writers of the text may be humans, natural language generation systems, or a combination of the two. Indeed, it is common to pit NLG systems against human writers. Clearly, to guard against any biases the judges might have, they must be unaware of which text is generated by which model. Having a panel of judges combats (but does not eliminate) the inherent subjectivity of human judgement of natural language. The rationale is that, although multiple judges rarely reach a consensus, their collective opinion can provide persuasive evidence about significant differences between different models. The main limitation of this approach is that it requires the specification of the texts to be evaluated to be simple enough to be easily articulated to the judges. This was the case in [40], where judges were told that each text was meant to be a general explanation of a given biological entity or process aimed at freshman biology students. For applications in which the input to the generation process must include a detailed characterization of the context (e.g., interactive applications in which output must be tailored to a complex user model or a history of previous interaction), it can be extremely difficult for the judges to fictitiously place themselves in the specific context in order to judge the texts.[11] As we have seen in the previous section, the input to GEA is complex. It consists of a possibly complex and novel argument subject (e.g., a new house with a

[11] See [6] for an illustration of how the specification of the context can become extremely complex when human judges are used to evaluate content selection strategies for a dialogue system.

large number of features), and a complex model of the user's preferences. Therefore, the human judges method is not deemed appropriate to evaluate GEA.

The *corpus-based* evaluation method can be applied when a corpus of input/output pairs is available [53]. The input consists of all relevant knowledge sources and the output can be either textual, graphical or multimedia. A portion of the corpus (the training set) is used to develop a computational model capable of generating the output from the corresponding input. The remainder of the corpus (the testing set) is used to evaluate the model. The model is evaluated by verifying each pair in the testing set to determine whether the output produced by the model applied to the input matches the corresponding output in the pair. The main advantage of the corpus-based method is that it does not require subjective human judgements about the quality of the generated text. However, it requires a large corpus of input/output pairs. In many cases, there is no extant corpus, and as the complexity of the input specification increases, the effort that would be required to create such a corpus becomes prohibitive. As discussed above, GEA makes use of a rich input representation, which includes an argument subject with many features, and a complex user model, and hence it was not possible to create such a corpus. Indeed, it is not even clear how one would go about it without using a generator such as GEA.

Arguably, all natural language processing tools are components of larger systems that are designed to assist users engaged in tasks, and therefore a natural and extremely informative way to evaluate their effectiveness is by experimenting with users performing those tasks [33]. This evaluation method is called *task efficacy*. As an early example of task efficacy evaluation in NLG consider [60]. In this study, an explanation generation model for a medical belief network was proven to be effective by showing that the explanations it generated improved diagnostic accuracy and increased user confidence in the final diagnosis. In general, the task efficacy evaluation method allows one to evaluate a generation model not by explicitly evaluating its output, but rather by measuring the output's effects on users' behaviors, beliefs and attitudes in the context of the task. The only requirement for this method is the specification of a sensible task. However, since task efficacy, to achieve sufficient statistical power, typically requires involving a large number of users in the evaluation, it is often considered the most expensive and difficult-to-organize method. Nevertheless, because the applicability of other methods is rather limited, task efficacy is becoming a common choice in NLG research [15,18,51,64], and forms the basis for the evaluation framework we have developed.

3.3. The evaluation framework

3.3.1. The user task

Generally speaking, a suitable task for evaluating evaluative arguments should be one in which (a) the user is required to perform evaluations and comparisons of objects in order to complete the task; and (b) presenting the user with evaluative arguments in the context of the task may change how the user performs the task along measurable dimensions.

A rather *basic* and *frequent* task satisfying these requirements is preferential choice: a selection task that has been extensively studied in decision analysis. It consists of having the decision-maker select a subset of preferred objects (e.g., houses) out of a set of possible alternatives by considering trade-offs among multiple objectives (e.g., house location, house quality) and by evaluating the objects with respect to their values for a set of primitive attributes (e.g., distance from work, size of the garden). The task we have devised for the evaluation framework is an extension of preferential choice and comprises two subtasks. As shown in Fig. 8, at the start of the first subtask the user is presented with information about a set of alternatives. Next, she is asked to select a subset of *n* preferred alternatives and to order them by preference in what is called a "Hot List". In the second subtask the user is presented with an evaluative argument about a new instance (not included in the initial set of alternatives), and she is asked whether she wants to include it in her Hot List. If the user's answer is affirmative, she has to decide where to place the new instance in the ordered Hot List. Finally, the user fills out a questionnaire about her attitudes and beliefs about the new instance and the decision task.

3.3.2. The data exploration environment

In our evaluation framework, the user performs this task by using a system for interactive data exploration and analysis (IDEA) [54]. The IDEA environment facilitates the user's autonomous exploration of the set of alternatives and the performance of the two subtasks in the real-estate domain. We chose this domain because it is familiar to almost everyone, but still presents a challenging decision task. The design of the IDEA system was inspired by the

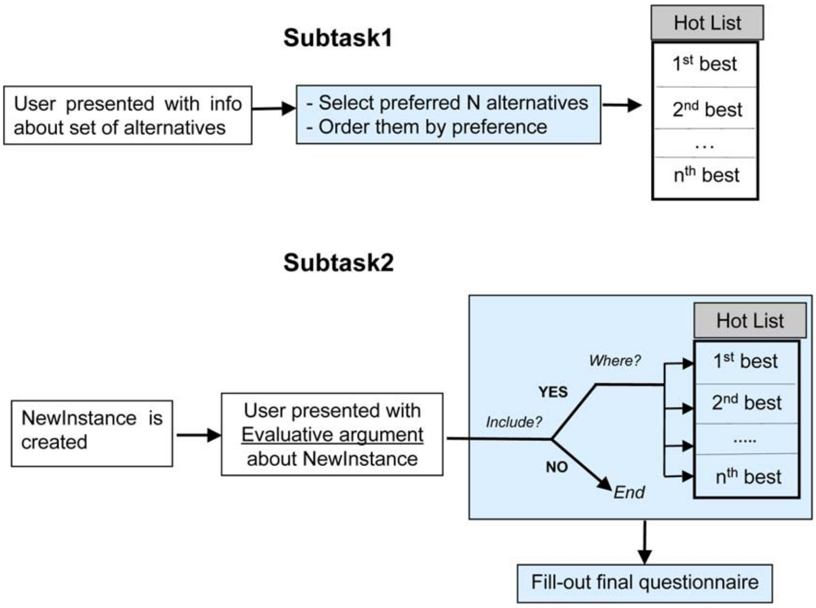

Subtask1
User presented with info about set of alternatives
- Select preferred N alternatives
- Order them by preference
Hot List
1st best
2nd best
...
nth best
Subtask2
NewInstance is created
User presented with Evaluative argument about NewInstance
Include?
YES
NO
Where?
End
Hot List
1st best
2nd best
.....
nth best
Fill-out final questionnaire

Fig. 8. User task at the core of the evaluation framework.

HomeFinder prototype [63] and details of the interface were refined by iterative evaluation with HCI experts and pilot subjects. The interface is shown in Fig. 9 (the reader should ignore the argument about NewHouse-3-26 for now). The user can visually inspect many aspects of the set of houses in the map, the bar charts and the table (14 distinct attributes, including 1 attribute of the street on which the house is located, and 2 attributes of the neighborhood). Furthermore, the user can explore this information by applying powerful interactive techniques including dynamic-queries, drag-and-drop and painting [54].

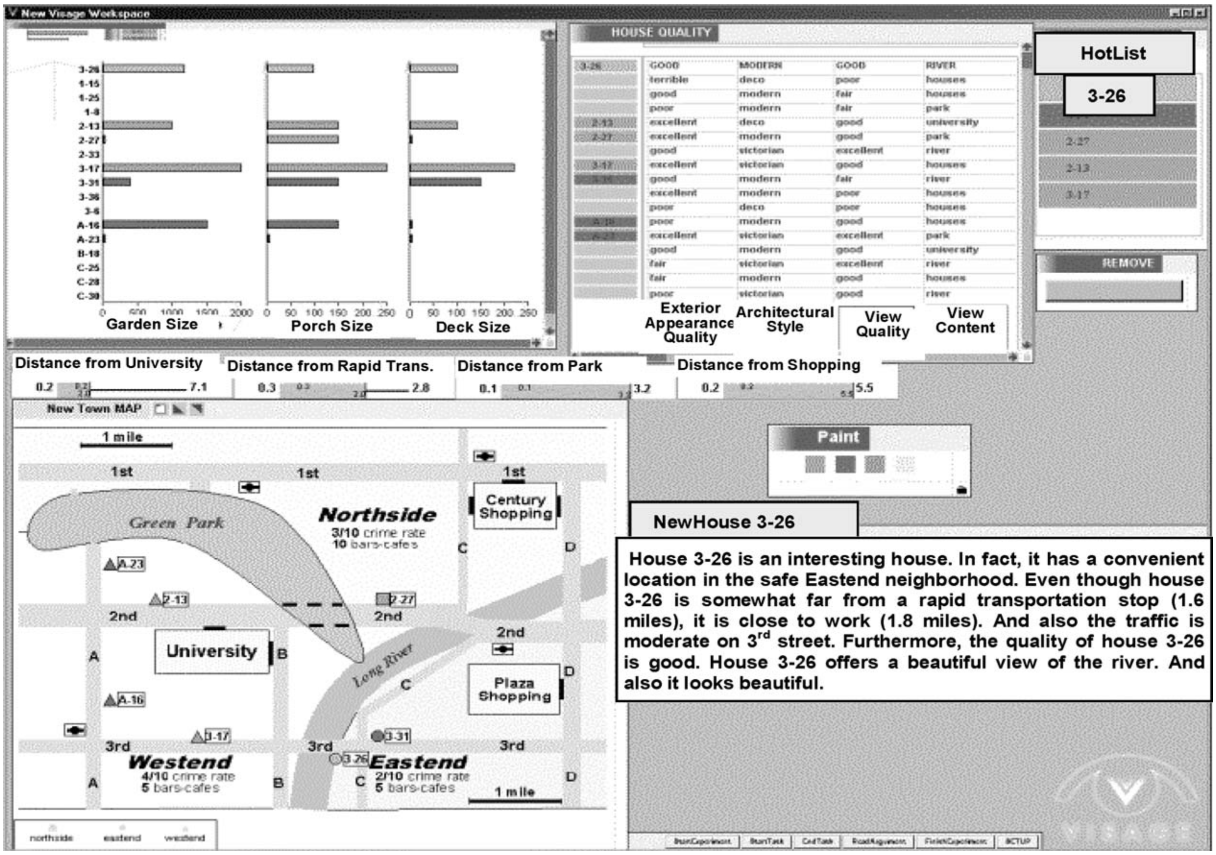
3.3.3. *The evaluation framework*

Fig. 10 illustrates the architecture of the evaluation framework, which consists of four main sub-systems: the IDEA system, the User Model Refiner, the New Instance Generator and GEA. The framework assumes that a model of the user's preferences (an AMVF) has been previously acquired from the user, to assure a reliable initial model. The user is assigned the task of selecting from the dataset the four most preferred alternatives by placing them in a "Hot List" (see Fig. 9, upper right corner) ordered by preference. When the user feels that the task is accomplished, the ordered list of preferred alternatives is saved as her Preliminary Hot List (Fig. 10(2)). Then, the User Model Refiner refines the initial model, making any adjustments necessary to make the model consistent with the preferences that the user expressed by creating her Hot List (Fig. 10(3)). This refinement process produces a Refined Model of the User's Preferences by heuristically adjusting the model weights. Then a New Instance (NewI) is designed on the fly by the New Instance Generator to be preferable for the user given her refined preference model (Fig. 10(4)). More precisely, the new instance is designed so that its value for the user is the average of the values of the two instances with the highest values in the HotList.

At this point, the stage is set for argument generation. Given the Refined Model of the User's Preferences, the Argument Generator produces an evaluative argument about NewI tailored to the model (Fig. 10(5)), which is presented to the user by the IDEA system (Fig. 10(6)) (see also Fig. 9 for an example). The argument goal is to persuade the user that NewI is worth being considered. Notice that all the information about NewI is also presented graphically.

Once the argument is presented, the user may (a) decide to immediately introduce NewI into her Hot List, (b) decide to further explore the dataset, possibly making changes and adding NewI to the Hot List, or (c) do nothing. Fig. 9 shows the display at the end of the interaction, when the user, after reading the argument, has decided to introduce NewI into the Hot List in first position (Fig. 9, top right).

When the user decides to stop exploring, and can thus be assumed to be satisfied with the selections in the hot list, measures related to the argument's effectiveness can be assessed (Fig. 10(7)).



	Exterior Appearance Quality	Architectural Style	View Quality	View Content
3-26	GOOD	MODERN	GOOD	RIVER
	terrible	deco	poor	houses
	good	modern	fair	houses
	poor	modern	fair	park
2-13	excellent	deco	good	university
2-27	excellent	modern	good	park
	good	victorian	excellent	river
3-17	excellent	victorian	good	houses
[illegible]	good	modern	fair	river
	excellent	modern	poor	houses
	poor	deco	poor	houses
[illegible]	poor	modern	good	houses
[illegible]	excellent	victorian	excellent	park
	good	modern	good	university
	fair	victorian	excellent	river
	fair	modern	good	houses
	poor	victorian	good	river

House 3-26 is an interesting house. In fact, it has a convenient location in the safe Eastend neighborhood. Even though house 3-26 is somewhat far from a rapid transportation stop (1.6 miles), it is close to work (1.8 miles). And also the traffic is moderate on 3rd street. Furthermore, the quality of house 3-26 is good. House 3-26 offers a beautiful view of the river. And also it looks beautiful.

Fig. 9. The IDEA environment display (at the end of the interaction).

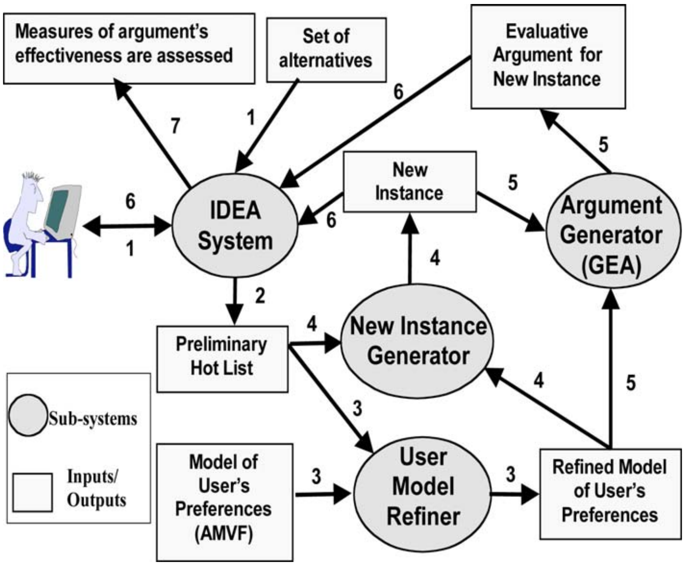


Fig. 10. Architecture of the evaluation framework.

(a) How would you judge the houses in your Hot List?
1st house
bad choice : ___ : ___ : ___ : ___ : ___ : ___ : ___ : _X_ : ___ : *good choice*
2nd house (New house) }
bad choice : ___ : ___ : ___ : ___ : ___ : ___ : _X_ : ___ : ___ : *good choice*
3rd house
bad choice : ___ : ___ : ___ : ___ : ___ : ___ : _X_ : ___ : ___ : *good choice*
4th house
bad choice : ___ : ___ : ___ : ___ : _X_ : ___ : ___ : ___ : ___ : *good choice*

Fig. 11. Sample self-report of user's satisfaction with houses in Hot List1.

3.3.4. Measures of argument effectiveness

Measures of argument effectiveness are obtained either from the record of the user's interaction with the system or from user self-reports in a final questionnaire (see Fig. 7 for an example of self-report) and include:

- Measures of behavioral intentions and attitude change: (a) whether or not the user adopts NewI; (b) in which position in the Hot List she places it; (c) how much she likes NewI and the other objects in the Hot List. Measures (a) and (b) are obtained from the record of the user interaction with the system, whereas measures in (c) are obtained from user self-reports.
- A measure of the user's confidence that she has selected the best for her in the set of alternatives. This measure and the ones below are all obtained from user self-reports.
- A measure of argument effectiveness derived by explicitly questioning the user at the end of the interaction about the rationale for her decision [48]. This can provide valuable information about which aspects of the argument were most influential in the user's decision making.
- Additional measures of argument effectiveness are derived by explicitly asking the user at the end of the interaction to judge the argument with respect to several dimensions of quality, such as content, organization, writing style and convincingness. However, evaluations based on judgements along these dimensions are clearly weaker than evaluations measuring actual behavioral and attitudinal changes [48].

A closer analysis of the measures of behavioral intentions and attitude change indicates that the measures in (c) are simply a more precise version of measures (a) and (b). In fact, not only do they assess, like (a) and (b), a preference ranking among the new alternative and the other objects in the Hot List, but they also offer two additional critical advantages: (i) Self-reports allow a subject to express differences in satisfaction more precisely than by ranking. For instance, in the self-report shown in Fig. 11, the subject was able to specify that the first house in the Hot List was only one unit of satisfaction better than the house following it in the ranking, while the third house was two unit better than the house following it. (ii) Self-reports do not force subjects to express a total order between the houses. For instance, in Fig. 11 the subject was allowed to express that the second and third houses in the Hot List were equally good for her.

Furthermore, measures of satisfaction obtained through self-reports can be combined in a single, statistically sound measure that concisely expresses how much the subject liked the new house with respect to the other houses in the Hot List. This measure is the z-score of the subject's self-reported satisfaction with the new house, with respect to the self-reported satisfaction with the houses in the Hot List. The z-score for an item, x_i indicates how far and in what direction, that item deviates from the mean of a population X, measured in units of the distribution's standard deviation.

Formally:

$$x_i \in X; \quad z(x_i) = \bigl(x_i - \mu(X)\bigr)/\sigma(X).$$

For instance, the satisfaction z-score for the new instance, given the sample self-reports shown in Fig. 11, would be: $[7 - \mu(8,7,7,5)]/\sigma(8,7,7,5) = 0.2$. Note that the satisfaction z-score for the new instance ranges in the $[-1.5, +1.5]$ interval. It is equal to the min value -1.5 when the new instance is scored 0 while all the other instances are scored 9, it is equal to the max value $+1.5$ when the new instance is scored 9 while all the other instances are scored 0, and it is equal to the mid point 0 when all the instances (including the new one) are equally scored.

Z scores, sometimes called "standard scores", are especially useful when comparing the relative standings of items from distributions with different means and/or different standard deviations. The satisfaction z-score precisely and concisely integrates all the measures of behavioral intentions and attitude change. We have used satisfaction z-scores as our primary measure of argument effectiveness.

To summarize, our evaluation framework supports users in performing a realistic task by interacting with an IDEA system. In the context of this task, an evaluative argument is generated by GEA and measurements of the argument's effectiveness are collected.

4. The experiment

In the previous section, we proposed a task-based framework for evaluating evaluative arguments. In the context of this task, an evaluative argument is generated by GEA and measurements are collected on its effectiveness. In this section, we report the results of an experiment run within this framework.

The design and development of GEA is based on several assumptions about what is necessary in order to generate effective evaluative arguments:

(1) Supporting and opposing evidence for the main evaluative claim should be identified and arranged according to a model of the reader's values and preferences. In GEA, it is assumed that such a model can be effectively represented as an AMVF, a quantitative model of preferences originally developed in decision theory.
(2) Evaluative arguments should be concise, presenting only pertinent and cogent information.
(3) Supporting and opposing evidence for the main evaluative claim should be carefully arranged according to the argumentative strategy presented in Section 2.2.3.
(4) The microplanning tasks, in their specific instantiation as described in Section 2.3, contribute to argument effectiveness.

Naturally, all these assumptions can be questioned and should be empirically tested. With respect to the first assumption, the question is whether a user specific AMVF is an effective model for tailoring evaluative arguments. As for the second assumption, nobody disputes that an argument, for the sake of brevity, should present only pertinent and cogent information. However, it remains an open question what the most effective degree of conciseness is. Concerning the third assumption, the argumentative strategy presented in Section 2.2.3 implements a set of guidelines from argumentation theory. However, alternative strategies could be more effective in specific situations or for a particular class of users. Finally, with respect to the fourth assumption, although it is generally accepted that some form of microplanning is needed to produce effective text, it is conceivable that implementations of the microplanning tasks other than the ones devised for GEA could be more effective.

The experiment we performed focuses on the empirical questions related to the first two assumptions; namely, whether a user specific AMVF is an effective model for tailoring evaluative arguments and what is the most effective degree of conciseness for evaluative arguments. To test the first assumption, we have compared the effectiveness of arguments tailored to the user's AMVF with the effectiveness of arguments tailored to a default AMVF, for whom all aspects of a house are equally important (i.e., all the weights in the AMVF are equal). To test the second assumption, as a preliminary attempt to determine an optimal level of conciseness for evaluative arguments, we have compared the effectiveness of arguments generated by our argument generator at two different levels of conciseness (i.e., for two values of the constant k, which controls argument conciseness in GEA as discussed in Section 2.2.3).

4.1. Experiment design and procedure

To address these questions, we designed a between-subjects experiment with four experimental conditions:

No-Argument: This is the baseline condition. In this condition, once subjects have completed the selection task, they are simply informed that a new house came on the market. No evaluative argument about the new house is presented to the subject. Information about the new house is only presented graphically.

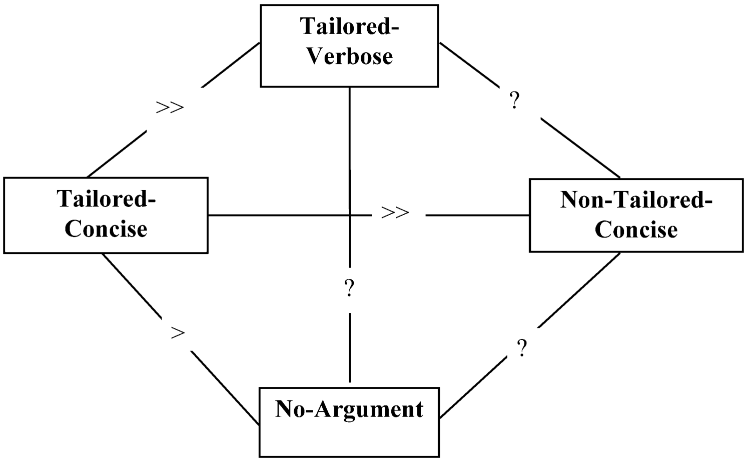


Fig. 12. Hypotheses on experiment outcomes.

Tailored-Concise: In this condition, once subjects have completed the selection task, they are presented with an evaluative argument about the new house tailored to their preferences and at a level of conciseness that we hypothesize to be optimal. Our assumption is that in our domain an effective argument should contain slightly more than half of the available evidence. By running the generator with different values for k on the user models of the pilot subjects, we found that this corresponds to $k = -0.3$. In fact, with $k = -0.3$ the arguments contained on average 10 pieces of evidence out of the 19 available (the AMVF contains 19 objectives).

Non-Tailored-Concise: In this condition, once subjects have completed the selection task, they are presented with an evaluation of the new house that, instead of being tailored to their preferences, is tailored to the preferences of a default average user, for whom all aspects of a house are equally important (i.e., all weights in the AMVF are the same). A similar default preference model is used for comparative purposes in [57]. The level of conciseness is still the one we hypothesize to be optimal (i.e., $k = -0.3$).

Tailored-Verbose: In this condition, once subjects have completed the selection task, they are presented with an evaluation of the new house tailored to their preferences, but at a level of conciseness that we hypothesize to be too low. We chose ($k = -1$), which in our analysis of the pilot subjects, corresponded on average to 16 pieces of evidence out of the possible 19.

In the four conditions, all the information about the new house is also presented graphically, so that no information is hidden from the user (see the new house House-3-26 in Fig. 9 for an example). And once the new house is introduced, subjects are free to perform data exploration to see how it compares to their Hot List choices.

Our hypotheses about the outcomes of the experiment are summarized in Fig. 12. We expect arguments generated for the Tailored-Concise (TC) condition to be more effective than arguments generated for both the Non-Tailored-Concise (NTC) and Tailored-Verbose (TV) conditions. We also expect the TC condition to be somewhat better than the No-Argument (NA) condition, but to a lesser extent, because subjects, in the absence of any argument, may spend more time further exploring the dataset, therefore reaching a more informed and balanced decision. Finally, we do not have strong hypotheses about comparisons of argument effectiveness among the No-Argument, Non-Tailored-Concise and Tailored-Verbose conditions.

The experimental procedure is summarized in Fig. 13. It consists of two phases. In the first phase, the subject fills out three online questionnaires. One questionnaire implements the SMARTER elicitation method from decision theory (see Section 2.2.2) to effectively acquire an AMVF model of the subject's preferences [21]. In our experiment, we can safely assume that the user preferences can be represented as an AMVF because there are no uncertain aspects in the user selection task.

The other two questionnaires assess the subject's argumentativeness (tendency to argue) [30], and need for cognition (tendency to engage in and to enjoy effortful cognitive endeavours) [3]. These are two key individual features that research in persuasion has shown to influence people's reaction to arguments [45]. Any experiment in persuasion

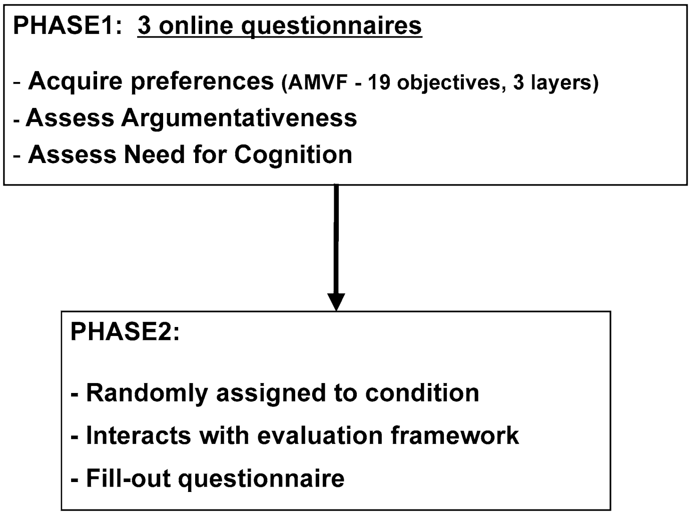

PHASE1: 3 online questionnaires
- Acquire preferences (AMVF - 19 objectives, 3 layers)
- Assess Argumentativeness
- Assess Need for Cognition
PHASE2:
- Randomly assigned to condition
- Interacts with evaluation framework
- Fill-out questionnaire

Fig. 13. Experimental procedure.

should control for these variables. In the second phase of the experimental procedure (see Fig. 13), to control for other possible confounding variables (including intelligence and self-esteem), the subject is randomly assigned to one of the experimental conditions. Then, the subject interacts with the evaluation framework, and at the end of the interaction the subject fills out a final questionnaire in which measures of the argument effectiveness are collected (see Section 2.3.4).

The experiment was first performed with 8 pilot subjects to refine and improve the experimental procedure. The instructions, the final questionnaire, and the script followed by the experimenter in presenting the IDEA system were checked for clarity.[12] Once the experimental procedure was sufficiently stable, we ran a formal experiment involving 40 subjects, 10 in each experimental condition. Each subject had only one interactive section with the evaluation framework.

4.2. *Experiment results*

During the analysis of the experimental outcomes, the data for four subjects were eliminated. One subject was eliminated from the TC condition because s/he was an outlier for the self-report measure indicating the subject's confidence in her decision process (the score was more than 3 standard deviations below the average). Another subject was eliminated from the NTC condition because s/he was an outlier for the "need for cognition" measure (the score was more than 2.5 standard deviations below the average); and also this subject's self-reports of satisfaction with the instances in the HotList were inconsistent with their explicit ranking (in the HotList). Finally, two more subjects were eliminated because their satisfaction self-report score for all the instances in the HotList and for the new instance was the maximum possible (i.e., 9 on the 1–9 scale). We consider this as an extremely anomalous situation for two reasons. First, and most importantly, because, as all the instances in the HotList were scored 9, the new instance had no chance to obtain a positive z-score (i.e., we are facing a kind of ceiling effect). Secondly, the fact that these two subjects gave all five instances the top score indicates that they have rather nondiscriminatory preferences. One of these two subjects was in the TC condition, the other in the NA condition.

4.2.1. *Effectiveness comparisons*

As discussed in Section 3.3.4, by precisely and concisely integrating all the measures of behavioral intentions and attitude change, the satisfaction z-score is the primary measure of argument effectiveness available in the framework. We focus on this measure first. Our statistical analysis was based on the Dennett test. This is the appropriate test for

[12] The three initial questionnaires are standard ones developed in decision theory and social psychology.

Table 1
Results for satisfaction z-scores when Tailored-Concise is compared with the other three conditions directionally

Cond-1	Cond-2	Mean difference (Cond-1 – Cond-2)	Std. error	Significance
NA (0.273)	TC (0.994)	−0.720	0.389	**0.086**
NTC (0.275)	TC (0.994)	−0.719	0.389	**0.087**
TV (0.047)	TC (0.994)	−0.947	0.380	**0.023**

Table 2
From Logs of the interaction: average time spent by subjects in the four conditions further exploring the dataset after the new house is presented. The difference between No-Argument and Tailored-Concise is significant

Condition	Time
NA	0:03:56
NTC	0:03:37
TC	0:02:44
TV	0:03:30

an experiment, like ours, in which there are several groups and the apriori goal is to compare one of them (i.e., TC) with each of the others [20].[13]

As shown in Table 1, the satisfaction z-scores obtained in the experiment provide support for our hypotheses. Arguments generated for the TC condition had greater satisfaction z-scores than arguments generated for the TV, NTC and NA conditions. The difference in effectiveness between arguments generated in the TC condition and arguments generated in the TV condition was statistically significant ($p < 0.05$), while the difference in the other two comparisons TC vs. NTC and NA was only marginally significant ($p < 0.1$). A possible reason/explanation why these differences were only marginally significant is our relatively small number of subjects. Another possibility is that the difference between our uniform default model (for the NTC) and the user specific one (for the TC) was too small. As we will see in Section 5.2, both possibilities are corroborated by a more recent study.

Remarkably, the TC appears to be better than the NA condition to a greater extent than we expected. We believed that in the absence of any argument, NA subjects would have spent more time further exploring the dataset, therefore reaching a more informed and balanced decision (with a satisfaction between TC and TV/NTC). However, although NA subjects did spend significantly more time further exploring the dataset (see Table 2), this was not enough to compensate for the lack of an explicit argument.

With respect to the other measures of argument effectiveness that we have considered (i.e., decision confidence, decision rationale and argument quality), we did not find any significant results.

4.2.2. Possible confounding variables

The design of the experiment takes into account the fact that the effectiveness of an argument is determined not only by the argument itself, but also by the subjects' traits such as argumentativeness (Arg), need for cognition (NFC), self-esteem and intelligence. The reason subjects are randomly assigned to one of the four conditions is precisely to control for these (and other) possible confounding variables.

As an extra check, the subjects' Arg and NFC were assessed before running the experiment in order to verify whether subjects had been successfully randomized to obtain four conditions with equivalent Arg and NFC. At first glance, the data in Table 3, reporting the means of Arg and NFC for each condition, seem to indicate that randomization failed, as TC was both the condition with the lowest Arg and the highest NFC (see Table 3). However, when we consider the differences between the means, it should be noted that they are minimal with respect to the ranges on which Arg and NFG can vary: $(-54, +54)$ and $(-50, +50)$ respectively. More tellingly, the means for Arg and NFC are all in the first half of the positive side of the ranges (i.e., moderately high). Since all results in social psychology

[13] In previous papers (e.g., [8]) we reported results that were based on applying the t-test in each pairwise comparison. However, we subsequently realized that the Dunnett test was more appropriate given our experimental design. Furthermore, our prior analysis included two subjects who have been excluded from the analysis we are reporting in this paper.

Table 3
Means of Argumentativeness (Arg) and Need for Cognition (NFC) for the four experimental conditions

Condition	Arg mean	NFC mean
NA	4.7	18.7
NTC	10.9	15
TC	**3.8**	**24.9**
TV	10.7	15

on the influence of Arg and NFC on persuasion have considered differences between individuals who scored *high* vs. *low* on these personality traits, we can assume with confidence that Arg and NFC did not substantially influence the outcome of our experiments.

Unfortunately, because of the limited number of subjects it was not possible to consider all relevant independent variables (i.e., argument-type, Arg and NFC) in a single generalized linear model.

5. Related work

In previous sections we have discussed related work whenever it was necessary as background for our research. In this section, we complete our analysis of previous work by focusing on two key aspects that require a more extensive treatment. First, we examine previous work on content selection and organization for evaluative arguments. Second, we review related work that either co-occurred or followed research on GEA. In particular, we consider projects that have extended GEA's approach and/or evaluated generators of user-tailored evaluative arguments.

5.1. Previous work on content selection and organization for evaluative arguments

Although considerable research has been devoted to content selection and organization for generating evaluative arguments, all approaches proposed were limited both in the type of evaluative arguments generated, and in the extent to which they comply with guidelines from the argumentation literature.

Ref. [47] describes a system that uses a measure of evidence strength to tailor evaluations of hotel rooms to its users. However, this system adopts a qualitative measure of evidence strength (an ordinal scale that appears to range from very-important to not-important). This limits the ability of the system to select and arrange argument evidence, because qualitative measures only support approximate comparisons and are notoriously difficult to combine (e.g., how many "somewhat-important" pieces of evidence are equivalent to an "important" piece of evidence?).

Refs. [6,13] studied the generation of evaluative arguments in the context of collaborative planning dialogues. Although they also adopt a qualitative measure of evidence strength, when an evaluation is needed this measure is mapped into numerical values so that preferences can be compared and combined more effectively. However, with respect to GEA this work makes two strong simplifying assumptions. It only considers the decomposition of the preference for an entity into preferences for its primitive attributes (not considering that complex preferences frequently have a hierarchical structure). Additionally, it assumes that the same dialogue turn cannot provide both supporting and opposing evidence.

The system described in [7] also employs additive decision models in recommending courses, though the focus of this work is on dynamically acquiring a model of the student's preferences. The system's recommendations are limited to recommending a single option that is considered better than the user's current choice. In addition, this work only addresses the problem of selecting positive attributes to justify the recommendation, and does not consider how to plan and realize the positive and negative attributes of multiple suggested options.

In [38], Kolln proposes a framework for generating evaluative arguments which is based on a quantitative measure of evidence strength. Evidence strength is computed on a fuzzy hierarchical representation of user preferences. Although this fuzzy representation may represent a viable alternative to the AMVF we have discussed in this paper, Kolln's proposal is rather sketchy in describing how his measure of strength can be used to select and arrange the argument content.

Finally, [35] is the previous work most relevant to our proposal. As described in Section 2.2.3, from this work we have adapted a measure of evidence strength (i.e., *compellingness*), and a measure that defines when a piece of evidence is worth mentioning (i.e., *notably-compelling?*). However, there are two key differences between Klein's

Table 4
Contributions of proposed argumentation strategy in context of previous work

	Quantitative measure of importance	Based on argumentation theory	Argument type	
			Single entity	Comparison
[47]	no	no	no	yes
[35]	yes	no	no	yes
[38]	yes	no	yes	no
[6,7]	yes	no	yes	no
[**Our Strategy**]	yes	yes	yes	yes

work and ours. Klein only developed strategies for comparison, and these strategies were not based on argumentation theory.

Table 1 summarizes the contributions of our proposal with respect to previous work on content selection and organization for generating evaluative arguments. The table considers three dimensions: whether the proposed approach uses a quantitative vs. a qualitative measure of evidence importance, whether the proposed approach is based on guidelines from argumentation theory, and whether the approach covers arguments evaluating a single entity or comparing two entities. It is clear that our strategy extends previous work in two ways: by covering both arguments evaluating a single entity, as well as arguments comparing two entities, and by implementing a comprehensive set of guidelines from argumentation theory

5.2. Related recent work that co-occurred with or followed our research

Several recent projects have extended GEA's approach and/or evaluated generators of user-tailored evaluative arguments. Overall, evidence from these studies indicates that tailoring an evaluative argument to a user-specific AMVF does increase its effectiveness. These studies seem to indicate that this hypothesis was only marginally confirmed in our experiment because we either did not run a sufficient number of subjects or the default model we considered for the NTC condition was not sufficiently different from the user-specific one.

STOP is a generator of user-tailored smoking cessation letters [51], where tailoring is based on information collected by means of a 4-page multiple choice questionnaire about the smoker's habits, health concerns and so forth. The STOP system is especially relevant to our research because one section of the generated letter is an evaluative argument. More specifically, the letter "motivation paragraph" mentions only user-specific "important" likes and dislikes about smoking (e.g., helps me to relax vs. it is expensive). The effectiveness of STOP has been tested in what is by far the most extensive, longest and costliest task-based evaluation of an NLG system: a clinical trial involving 2553 smokers. In this study, smokers were randomly assigned to three groups which respectively received a tailored letter, a non-tailored letter and no letter. Effectiveness of tailoring was tested six months later by asking smokers whether they had quit smoking or not (positive answers were checked through saliva samples). In general, the results of the STOP evaluation were inconclusive. Although it seems that tailored letters may have been better than non-tailored ones among smokers for whom quitting was especially difficult, the difference in effectiveness among the three conditions was not overall statistically significant. With respect to our work, this study does not provide any positive or negative evidence to the hypotheses we tested in our experiment. Although STOP generates text that is partially an evaluative argument tailored to the user, it does not follow an approach in which arguments are tailored to a user-specific AMVF. The aspect of STOP's evaluation most relevant to our experiment is the detailed analysis of why the evaluation failed to prove tailoring to be effective. Four possible reasons were considered: (i) tailoring cannot have much effect in inducing smoking cessation, receiving a letter is what matters not the actual content; (ii) tailoring should have been based on different or more-complex knowledge about the smokers; (iii) the knowledge of the users was appropriate but the tailoring was inappropriate; (vi) STOP tailoring has an effect, but only a larger clinical trial could show it. In practice, the outcome of the STOP evaluation cannot tell us which of these reason(s) is playing a role. Interestingly, the same could be said of the results of our experiment which showed that the difference between tailored and non-tailored arguments was only marginally significant. However, as we will see shortly, the outcome of a recent and more extensive evaluation of MATCH, a system similar to GEA, indicates that tailoring to an AMVF does have a significant effect and so the most likely reason for our marginally significant findings is (vi) i.e., an insufficient number of data-points.

The MATCH system [62] extends GEA's approach to generating evaluative arguments. MATCH is a multimodal, speech-enabled dialogue system implemented on a PDA and intended to help people find information about restaurants and subway routes in New York City. Empirical testing of spoken dialogue systems has shown that presentation of complex information resulting from a user request is one of the most time-consuming phases of a dialogue. One of the key research goals behind the MATCH project is to improve the information presentation phase by enabling the system to select only the most relevant information and to effectively present it. To achieve this goal, MATCH adopts GEA's decision-theoretic framework in which user preferences are modelled as AMVFs. Like GEA, MATCH relies on the user-specific AMVF to generate cogent, concise text whose content and organization is tailored to the user. The system can generate three different types of user-tailored presentations: recommendations, comparisons and summaries. A *recommendation* is an evaluative argument about the best available alternative. A *comparison* is an evaluative argument comparing at most five alternatives pointing out reasons for choosing each of them, while a *summary* simply provides an overview of a set of alternatives highlighting the attributes for which they are most (dis)similar. [62] evaluated the effectiveness of these three presentation types in a within-subjects experiment in which each participant "overheard" a series of dialogues about selecting a restaurant. In each session, the participant was presented with an argument generated according to one strategy and either tailored to her own model or to the model of another randomly selected participant (note that tailoring the argument to a randomly chosen user is a much more extreme choice than using a uniform default model as we did in GEA). At the end of each session the participant was asked to rate the information quality of the argument on a 0–5 scale. The experiment involved 16 participants. Because the experimental setting was based on overhearing conversations, it was possible to run a large number of sessions (64 in the actual experiment) for each participant for a total of 1024 sessions. Since in each session the participant is asked to express an information quality judgment on the proposed argument, 1024 information quality judgements were collected in the experiment. Note that this is a much larger number than the 36 judgements considered in our evaluation of GEA. To obtain the same number of judgments in our between-subjects evaluation framework we should have run 1024 participants! The results of this study provide further empirical evidence for the first of the two hypotheses we tested in our experiments; namely, that tailoring evaluative arguments to a user-specific AMVF increases their effectiveness. A two-way ANOVA for information quality by strategy and model only showed a significant effect for strategy ($F = 127.9$, $p = 0.0001$), with summaries being clearly the least effective (summaries scored 2.33, comparisons 3.53 and recommendations 4.08). However, when the distance[14] between the randomly selected user model and the participant user model was considered, a paired t-test for information quality by user model over all strategies was highly significant. In particular when the distance was greater than 0.2 (which left a set of 464 paired comparisons), tailored presentations were preferred ($df = 463$, $t = 2.61$, $p = 0.009$). Because the average distance between user models in this study was 0.57, this result indicates that users are sensitive to relatively small perturbations of their models.

The FLIGHTS system [46] represents the most recent attempt to generate user-tailored evaluative arguments in spoken dialogue. Like MATCH, FLIGHTS concisely compares complex options (i.e., flights) pointing out the most relevant information for the intended user. However, FLIGHTS demonstrates that tailoring to user preferences must be carried out at all levels of information presentation, so that not only is appropriate content selected, but it is presented appropriately in the current dialogue context, and with intonation that expresses contrasts intelligibly [49]. FLIGHTS employs more sophisticated content planning strategies, capable of generating plans with a richer discourse structure including the distinctions between theme/rheme given/new and in/definite. This information structure can then support finer-grain choices in linguistic realization as well as intelligent control of prosody to convey meaning, following the theory presented in [58]. At the time of this writing, an evaluation of FLIGHTS is planned following the same experimental design that has been successfully used to test the MATCH system [62].

[14] The distance between two AMVFs is defined as the sum, over all attributes, of the absolute values of the difference between the rankings for each attribute.

6. Conclusions and future work

The research presented in this paper is very interdisciplinary. We have integrated and extended principles and techniques form argumentation theory, decision theory, computational linguistics, social psychology and human computer interaction.

Our research makes three key contributions. First, we have developed a complete computational model for generating evaluative arguments tailored to the user's preferences. Second, we have devised and implemented an evaluation framework in which the effectiveness of evaluative arguments can be measured with real users. Third, within the framework, we have performed an experiment to test that our proposal for tailoring an evaluative argument to the user's preferences increases its effectiveness, and that differences in conciseness significantly influence argument effectiveness. While the second hypothesis was confirmed in the experiment, the first one was only marginally confirmed. However, independent testing by other researchers has recently provided further support for this hypothesis.

A key goal of the research described in this paper was to complete the research cycle that begins with developing a computational model, devising techniques to evaluate the model, and applying these techniques to actually evaluate aspects of the model. To achieve this goal, and because of the complexity of the issues involved, we had to limit our investigation in several ways. Clearly, all limitations are open doors for future research.

More complex arguments: Many naturally occurring arguments consist of a mixture of evaluative arguments and other basic types of arguments (i.e., factual, causal and recommendation). Although the focus of this work has been on purely evaluative arguments, the long term goal of our research is to develop testable models for generating arguments that combine causal arguments, evaluative arguments and recommendations. We plan to integrate our work on generating evaluative arguments from an AMVF with previous work on generating causal arguments from Bayesian Networks [60,67], as well as previous work on generating recommendations from influence diagrams [32].

Automatic acquisition of linguistic knowledge: Another limitation of our model for generating evaluative arguments is that the human developer needs to encode most of the linguistic knowledge sources, which include the rhetorical strategies, the lexicon as well as the sentence-planning strategies (e.g., aggregation, generation of referring expressions). The problem with this human-intensive process is that it is extremely time-consuming, it has to be repeated for any new domain and most importantly the resulting knowledge tends to be brittle (i.e., its performance abruptly decreases when unexpected situations arise). To address this problem, we plan to supplement the intuition of the human developer with a probabilistic, data-driven procedure for the automatic acquisition of linguistic knowledge about evaluative arguments from text corpora [11,52].

Beyond AMVF: Although an AMVF can reasonably model most people's preferences in many situations, it does make strong assumptions of independence across attributes. So, in some settings, it might be necessary to use more complex models of preferences that take attribute interactions into account. Notice that adopting a different model of the user's preferences would require redefining the measure of evidence importance used in the argumentation strategy to select and order the content of the argument. Also, as an additional difficulty, there are no "simple and quick" methods from decision theory (similar to SMARTER [21]) to acquire these models. However, a set of novel techniques that apply machine learning to the problem of preference elicitation may help in this respect [10,28].

Generating multimedia arguments: As we have seen in Section 3.3, in the evaluation framework, the argument about the new house is presented in the context of a graphical display, which shows all the information about the new house (and all the information about the other houses). However, there is no integration between the argument and the display. In the current architecture, GEA generates the evaluative arguments without considering how the information is displayed graphically. A direction for future work is to study how natural language evaluative arguments can be integrated with graphics at an increasing level of sophistication:

(i) The information graphic is given and cannot be changed. However, the natural language generator, while planning the text, can access a representation of how the information is displayed graphically. So, the generated argument can be enhanced by adding references to the graphics by indicating, for instance, where in the visualization the user can find the information mentioned in the argument.

(ii) Once the textual argument is planned, graphics can also be enhanced to make the argument more effective. For instance, information mentioned in the argument can be highlighted in the graphics in a way that indicates its role in supporting or opposing the argument claims.

(iii) Graphics and text are planned together to achieve an abstract communicative goal. This sophisticated integration between text and graphics may require major architectural changes in GEA. However, we have recently made progress in this area, see [26].

References

[1] F.H. Barron, B.E. Barrett, Decision quality using ranked attribute weights, Management Science 42 (11) (1996) 1515–1523.

[2] J. Blythe, Visual exploration and incremental utility elicitation, in: Proceedings of the National Conference on Artificial Intelligence, 2002, pp. 526–532.

[3] J.T. Cacioppo, R.E. Petty, C.F. Kao, The efficient assessment of need for cognition, Journal of Personality Assessment 48 (3) (1984) 306–307.

[4] J.T. Cacioppo, R.E. Petty, K.J. Morris, Effects of need for cognition on message evaluation, recall, and persuasion, Journal of Personality and Social Psychology 45 (4) (1983) 805–818.

[5] C.B. Callaway, J.C. Lester, Narrative prose generation, Artificial Intelligence 139 (2) (2002) 213–252.

[6] S. Carberry, J. Chu-Carroll, Collaborative response generation in planning dialogues, Computational Linguistics 22 (2) (1998) 355–400.

[7] S. Carberry, J. Chu-Carroll, S. Elzer, Constructing and utilizing a model of user preferences in collaborative consultation dialogues, Computational Intelligence Journal 15 (3) (1999) 185–217.

[8] G. Carenini, J.D. Moore, An empirical study of the influence of user tailoring on evaluative argument effectiveness, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence, Seattle, USA, 2001, pp. 1307–1314.

[9] J. Chai, V. Horvath, N. Nicolov, M. Stys, N. Kambhatla, W. Zadrozny, P. Melville, Natural language assistant: A dialog system for online product recommendation, AI Magazine 23 (2) (2002) 63–75.

[10] U. Chajewska, D. Koller, D. Ormoneit, Learning an agent's utility function by observing behavior, in: Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 35–42.

[11] J. Chen, S. Bangalore, O. Rambow, M.A. Walker, Towards automatic generation of natural language generation systems, in: Proceedings of the 19th International Conference on Computational Linguistics (COLING), Taipei, Taiwan, 2002, pp. 1–7.

[12] C.-F. Chien, F. Sainfort, Evaluating the desirability of meals: An illustrative multiattribute decision analysis procedure to assess portfolios with interdependent items, Multi-Criteria Decision Analysis 7 (4) (1998) 230–238.

[13] J. Chu-Carroll, S. Carberry, A plan-based model for response generation in collaborative task-oriented dialogues, in: Proceedings of the National Conference on Artificial Intelligence, AAAI Press, Menlo Park, CA, 1994, pp. 799–805.

[14] R.T. Clemen, Making Hard Decisions: An Introduction to Decision Analysis, second ed., Duxbury Press, Belmont, CA, 1996.

[15] N. Colineau, C. Paris, K. Vander Linden, An evaluation of procedural instructional text, in: Proceedings International Natural Language Generation Conference, 2002, pp. 128–135.

[16] E.P.J. Corbett, R.J. Connors, Classical Rhetoric for the Modern Student, Oxford University Press, Oxford, 1999.

[17] V. Demberg, J.D. Moore, Information presentation in spoken dialogue systems, in: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006, pp. 65–72.

[18] B. Di Eugenio, M. Glass, M.J. Trolio, The DIAG experiments: Natural language generation for intelligent tutoring systems, in: Proceedings International Natural Language Generation Conference, 2002, pp. 120–127.

[19] B. Di Eugenio, J.D. Moore, M. Paolucci, Learning features that predict cue usage, in: Proceedings of the 35rd Annual Meeting of the Association for Computational Linguistics, 1997, pp. 80–87.

[20] C.W. Dunnett, A multiple comparison procedure for comparing several treatments with a control, Journal of the American Statistical Association 50 (1955) 1096–1121.

[21] W. Edwards, F.H. Barron, SMARTS and SMARTER: Improved simple methods for multiattribute utility measurements, Organizational Behavior and Human Decision Processes 60 (1994) 306–325.

[22] M. Elhadad, Using argumentation in text generation, Journal of Pragmatics 24 (1995) 189–220.

[23] M. Elhadad, K.R. McKeown, J. Robin, Floating constraints in lexical choice, Computational Linguistics 23 (2) (1997) 195–239.

[24] M. Elhadad, J. Robin, An overview of SURGE: A reusable comprehensive syntactic realization component, Technical Report 96-03, Department of Mathematics and Computer Science, Ben Gurion University, Beer Sheva, Israel, 1996.

[25] P.C. Gordon, B.J. Grosz, L.A. Gilliom, Prounouns, names and the centering of attention in discourse, Cognitive Science 17 (3) (1993) 311–348.

[26] N.L. Green, G. Carenini, S. Kerpedjiev, J. Mattis, J.D. Moore, S.F. Roth, Autobrief: An experimental system for the automatic generation of briefings in integrated text and information graphics, International Journal of Human-Computer Studies 61 (1) (2004) 32–70.

[27] B.J. Grosz, A.K. Joshi, S. Weinstein, Centering: A framework for modeling the local coherence of discourse, Computational Linguistics 21 (2) (1995) 203–226.

[28] V. Ha, P. Haddawy, Similarity of personal preferences: Theoretical foundations and empirical analysis, Artificial Intelligence 146 (2) (2003) 149–173.

[29] K.J. Hee, M. Glass, R. Freedman, M.W. Evens, Learning the use of discourse markers in tutorial dialogue for an intelligent tutoring system, in: Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society, Philadelphia, USA, 2000, pp. 262–267.

[30] D.A. Infante, A.S. Rancer, A conceptualization and measure of argumentativeness, Journal of Personality Assessment 46 (1982) 72–80.

[31] A. Jameson, R. Schafer, J. Simons, T. Weis, Adaptive provision of evaluation-oriented information: Tasks and techniques, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, 1995, pp. 1886–1895.

[32] H.B. Jimison, L.M. Fagan, D.R. Shacter, H.E. Shortliffe, Patient-specific explanation in models of chronic disease, Artificial Intelligence in Medicine 4 (1992) 191–205.

[33] K. Sparck Jones, Automatic language and information processing: Rethinking evaluation, Natural Language Engineering 7 (1) (2001) 29–46.

[34] R.L. Keeney, H. Raiffa, Decisions with Multiple Objectives: Preferences and Value Tradeoffs, John Wiley and Sons, New York, 1976.

[35] D.A. Klein, Decision Analytic Intelligent Systems: Automated Explanation and Knowledge Acquisition, Lawrence Erlbaum Associates, 1994.

[36] A. Knott, R. Dale, Using linguistic pheomena to motivate a set of coherence relations, Discourse Processes 18 (1) (1994) 35–62.

[37] A. Knott, C. Mellish, A feature-based account of the relations signalled by sentence and clause connectives, Language and Speech 39 (1996) 143–183.

[38] M.E. Kolln, Employing user attitudes in text planning, in: Proceedings of the 5th European Workshop on Natural Language Generation, Leiden, The Netherlands, 1995, pp. 163–179.

[39] D. Kudenko, M. Bauer, D. Dengler, Group decision making through mediated discussions, in: Proceedings of the User Modelling Conference, Johnstown, Pennsylvania, USA, August 2003, pp. 238–247.

[40] J. Lester, B. Porter, Developing and empirically evaluating robust explanation generators: The KNIGHT experiments, Computational Linguistics 23 (1) (1997) 65–101.

[41] D. Marcu, The conceptual and linguistic facets of persuasive arguments, in: ECAI Workshop—Gaps and Bridges: New Directions in Planning and Natural Language Generation, 1996, pp. 43–46.

[42] K.J. Mayberry, R.E. Golden, For Argument's Sake: A Guide to Writing Effective Arguments, Harper Collins, College Publisher, 1996.

[43] W.J. McGuire, The nature of attitudes and attitudes change, in: G. Lindzey, E. Aronson (Eds.), The Handbook of Social Psychology, vol. 3, Addison-Wesley, Reading, MA, 1968, pp. 136–314.

[44] W.J. McGuire, The nature of attitudes and attitudes change, in: G. Lindzey, E. Aronson (Eds.), Handbook of Social Psychology, vol. 3, second ed., Addison-Wesley, Reading, MA, 1969, pp. 136–314.

[45] M.D. Miller, T.R. Levine, Persuasion, in: An Integrated Approach to Communication Theory and Research, Lawrence Erlbaum Associates, 1996, pp. 261–276.

[46] J.D. Moore, M.E. Foster, O. Lemon, M. White, Generating tailored, comparative descriptions in spoken dialogue, in: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, AAAI Press, 2004, pp. 917–922.

[47] K. Morik, User models and conversational settings: Modeling the user's wants, in: A. Kobsa, W. Wahlster (Eds.), User Models in Dialog Systems, in: Symbolic Computation Series, Springer-Verlag, New York, 1989, pp. 364–385.

[48] J.M. Olso, M.P. Zanna, Attitudes and beliefs: Attitude change and attitude behavior consistency, in: R.M. Baron, W.G. Graziano (Eds.), Social Psychology, Holt, Rinehart and Winston, New York, 1991, pp. 192–225.

[49] S. Prevost, M. Steedman, Specifying intonation from context for speech synthesis, Speech Communication 15 (1994) 139–153.

[50] E. Reiter, R. Dale, Building Natural Language Generation Systems, Cambridge University Press, Cambridge, 2000.

[51] E. Reiter, R. Robertson, L.M. Osman, Lessons from a failure: Generating tailored smoking cessation letters, Artificial Intelligence 144 (1–2) (2003) 41–58.

[52] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, in: Proc. of the Conf. on Empirical Methods in NL Processing, Sapporo, Japan, 2003, pp. 105–112.

[53] J. Robin, K. McKeown, Empirically designing and evaluating a new revision-based model for summary generation, Artificial Intelligence 85 (1–2) (1996) 135–179.

[54] S.F. Roth, M.C. Chuah, S. Kerpedjiev, J.A. Kolojejchick, P. Lucas, Towards an information visualization workspace: Combining multiple means of expression, Human-Computer Interaction Journal 12 (1–2) (1997) 131–185.

[55] D. Scott, C. Sieckenius de Souza, Getting the message across in RST-based text generation, in: R. Dale, C. Mellish, M. Zock (Eds.), Current Research in Natural Language Generation, Academic Press, New York, 1990, pp. 47–73.

[56] M.R. Solomon, Consumer Behavior: Buying, Having, and Being, Prentice-Hall, Englewood Cliffs, NY, 1998.

[57] J. Srivastava, T. Connolly, L.R. Beach, Do ranks suffice? A comparison of alternative weighting approaches in value elicitation, Organizational Behavior and Human Decision Process 63 (1) (1995) 112–116.

[58] M. Steedman, Information-structural semantics for English intonation, in: M. Gordon, D. Büring, C. Lee (Eds.), LSA Summer Institute Workshop on Topic and Focus, Santa Barbara, July 2001, Kluwer Academic, Dordrecht, 2004, pp. 245–264.

[59] A. Stent, A conversation acts model for generating spoken dialogue contributions, Computer Speech and Language 16 (3) (2002) 313–352.

[60] H. J Suermondt, G.F. Cooper, An evaluation of explanations of probabilistic inference, Computers and Biomedical Research (1993) 242–254.

[61] K. Vander Linden, J.H. Martin, Expressing rhetorical relations in instructional text: A case study of the purpose relation, Computational Linguistics 21 (1) (1995) 29–58.

[62] M.A. Walker, S.J. Whittaker, A. Stent, P. Maloor, J.D. Moore, M. Johnston, G. Vasireddy, Generation and evaluation of user-tailored responses in multimodal dialogue, Cognitive Science 28 (2004) 811–840.

[63] C. Williamson, B. Shneiderman, The dynamic homefinder: Evaluating dynamic queries in a real-estate information exploration system, in: B. Shneiderman (Ed.), Sparks of Innovation in Human-Computer Interaction, Ablex Publishing Corp, ACM SIGIR, 1993, pp. 295–308.

[64] R.M. Young, Using Grice's maxim of quantity to select the content of plan descriptions, Artificial Intelligence 115 (2) (1999) 215–256.

[65] R.M. Young, J.D. Moore, DPOCL: A principled approach to discourse planning, in: Proceedings of the 7th International Workshop on Natural Language Generation, Kennebunkport, ME, June 17–21, 1994, pp. 13–20.

[66] R.M. Young, J.D. Moore, M.E. Pollack, Towards a principled representation for discourse plans, in: Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society, Lawrence Erlbaum Associates, Hillsdale, NJ, 1994, pp. 946–951.

[67] I. Zukerman, R. McConachy, K.B. Korb, Bayesian reasoning in an abductive mechanism for argument generation and analysis, in: Proc. AAAI Conference, 1998, pp. 833–838.