An Empirical Study of the Influence of User Tailoring on Evaluative Argument Effectiveness

Giuseppe Carenini
Department of Computer Science
University of British Columbia
Vancouver, B.C. Canada V6T 1Z4
carenini@cs.ubc.ca

Johanna D. Moore
The Human Communication Research Centre,
University of Edinburgh,
2 Buccleuch Place, Edinburgh EH8 9LW, UK.
jmoore@cogsci.ed.ac.uk

Abstract

The ability to generate effective evaluative arguments is critical for systems intended to advise and persuade their users. We have developed a system that generates evaluative arguments that are tailored to the user, properly arranged and concise. We have also devised an evaluation framework in which the effectiveness of evaluative arguments can be measured with real users. This paper presents the results of a formal experiment we performed in our framework to verify the influence of user tailoring on argument effectiveness.

1 Introduction

Evaluative arguments are pervasive in natural human communication. In countless situations, people attempt to advise or persuade their interlocutors that something is good (vs. bad) or right (vs. wrong). For instance, doctors need to advise their patients on which treatment is best for them (the patients). A teacher may need to convince a student that a certain course is (is not) the best choice for the student. And a sales person may need to compare two similar products and argue why her current customer should like one more than the other. With the explosion of the information available on-line and the ever-increasing availability of wireless devices, we are witnessing a proliferation of computer systems that aim to support or replace humans in similar communicative settings. Clearly, the success of these systems serving as personal assistants, advisors, or shopping assistants (e.g., [Chai, Budzikovaska et al. 2000]) may crucially depend on their ability to generate and present effective evaluative arguments.

In the last decade, considerable research has been devoted to develop computational models for automatically generating and presenting evaluative arguments. Several studies investigated the process of selecting and structuring the content of the argument (e.g., [Morik 1989; Elzer, Chu-Carroll et al. 1994; Klein 1994]), while [Elhadad, McKeown et al. 1997] developed a detailed model of how the selected content should be realised into natural language. All these approaches to evaluative argument generation follow a basic guideline from argumentation theory [Mayberry and Golden 1996]: effective evaluative arguments should be constructed considering the values and preferences of the audience towards the information presented. In practice, this means that all previous approaches tailor the generated arguments to a model of the user's values and preferences.

However, a key limitation of previous work is that none of the proposed approaches has been empirically evaluated. Thus, in particular, it is not clear whether and to what extent tailoring an evaluative argument to a model of the user increases its effectiveness. The work presented in this paper is a first step toward addressing this limitation. By recognising the fundamental role of empirical testing in assessing progress, generating new research questions and stimulating the acceptance of a technique as viable technology, we have performed an experiment to test the influence of user-tailoring on argument effectiveness. In the remainder of the paper, we first provide a short description of our system for generating evaluative arguments tailored to a model of the user's preferences. Then, we briefly present a framework to measure the effectiveness of evaluative arguments with real users. Next, we discuss the experiment we ran within the framework to test the influence of user tailoring on evaluative argument effectiveness.

2 User Tailored Evaluative Arguments

Our generation system, known as the Generator of Evaluative Argument (GEA) [Carenini 2000], generates evaluative arguments whose content, organisation and phrasing are tailored to a quantitative model of the user's values and preferences. The model is expressed as an Additive Multiattribute Value Function (AMVF), a conceptualization based on MultiAttribute Utility Theory (MAUT) [Clemen 1996]. Besides being widely used in decision theory (where they were originally developed), conceptualizations based on MAUT have recently become a common choice in the user modeling field [Jameson, Schafer et al. 1995]. Furthermore, similar models are also used in Psychology, in the study of consumer behaviour [Solomon 1998]. In GEA, a user specific AMVF is a key knowledge source in all the phases of the generation process. GEA is implemented as a standard

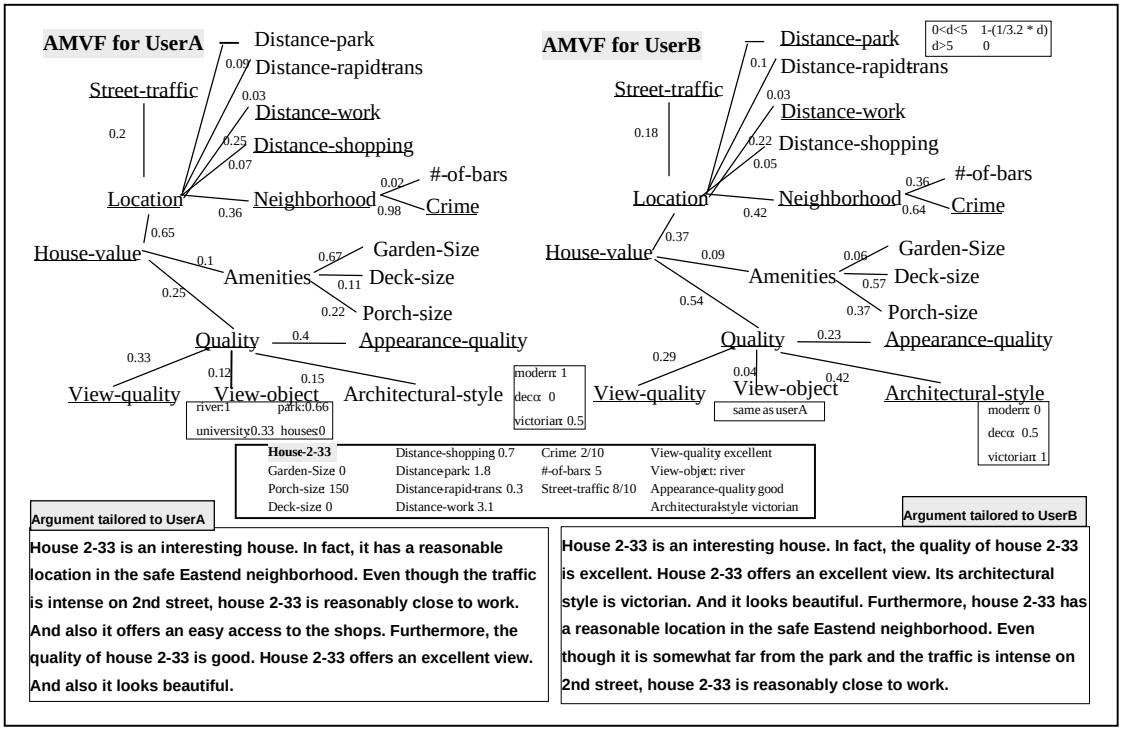

Figure 1 Top: AMVF for two sample users; for clarity's sake only a few component value functions are shown. Bottom: arguments about house-2-33 tailored to the two different models

pipelined generation system, including a text planner, a microplanner, and a sentence realizer.

2.1 AMVFs and their Use in GEA

An AMVF is a model of a person's values and preferences with respect to entities in a certain class. It comprises a *value tree* and a set of *component value functions*. A value tree is a decomposition of an entity value into a hierarchy of entity aspects (called objectives in decision theory), in which the leaves correspond to the entity primitive objectives (see top of Figure 1 for two simple value trees in the real estate domain). The arcs in the tree are weighted to represent the importance of an objective with respect to its siblings (e.g., in Figure 1 *location* for UserA is more than twice as important as *quality* in determining the *house-value*). The sum of the weights at each level is always equal to 1. A component value function for a primitive objective expresses the preferability of each value for that objective as a number in the [0,1] interval, with the most preferable value mapped to 1, and the least preferable one to 0. For instance, in Figure 1 the *victorian* value of the primitive objective *architectural-style* is the most preferred by UserB, and a *distance-from-park* of 1 mile has for UserB preferability (1 - (1/3.2 * 1))=0.69. Formally, an AMVF predicts the value $v(e)$ of an entity e as follows:

$v(e) = v(x_1,\ldots,x_n) = \Sigma w_i\, v_i(x_i)$, where

- $(x_1,\ldots,x_n)$ 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 $0 \leq w_i \leq 1$ and $\Sigma w_i = 1$; 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.

Thus, given someone's AMVF, it is possible to compute how valuable an entity is to that individual. Although for lack of space we cannot provide details here, given a user specific AMVF and an entity, GEA can also compute additional precise measures that are critical in generating a user-tailored evaluative argument for that entity. First, GEA can compute how valuable any objective of the entity is for that user. This information plays an essential role in phrasing the argument by determining the selection of scalar adjectives (e.g., convenient), which are the basic linguistic resources to express evaluations. Second, GEA can identify what objectives can be used as supporting or opposing evidence for an evaluative claim. Third, GEA can compute for each objective the strength of supporting (or opposing) evidence it can provide in determining the evaluation of its parent objective. In this way, in compliance with argumentation theory, evidence can be arranged according to its strength and concise arguments can be generated by only including sufficiently strong evidence [Carenini and Moore 2000]. The measure of evidence strength and the threshold that defines when a piece of evidence is worth mentioning were adapted from [Klein 1994].

A final note on AMVF's applicability. According to decision theory, in the general case, when uncertainty is

present, user's preferences for an entity can be represented as an AMVF only if her preferences for the primitive objectives satisfy a stringent condition (i.e., additive independence). However, evidence has shown that an AMVF is a reasonable model of most people's preferences under conditions of certainty [Clemen 1996]. We felt that we could safely use AMVFs in our study, because we selected the objectives to avoid possible violations of additive independence. And we considered a situation with no uncertainty.

2.2 An Example: Generating Arguments for Two Different Users

Figure 1 illustrates how the content, organization and phrasing of the arguments generated by GEA are sensitive to the model of a user's preferences. The top of the figure shows two different models of actual users in the real-estate domain. The bottom of the figure shows two evaluative arguments generated for the same house but tailored to the two different models. The primitive objectives' values for the house are reported in the middle of the figure. Notice how the two arguments differ substantially. Different objectives are included (the objectives included are underlined in the two models). Furthermore, the objectives are ordered differently (e.g., in the first argument *location* comes before *quality*, whereas the opposite is true in the second argument). Finally, the evaluations are also different. For instance, *quality* is *good* for UserA, but *excellent* for UserB.

3 The Evaluation Framework

To run our formal experiment, we used an evaluation framework based on the *task efficacy* evaluation method [Carenini 2000]. This method allows the experimenter to evaluate a generation model indirectly, by measuring the effects of its output on user's behaviors, beliefs and attitudes in the context of a task. Aiming at general results, we chose a basic and frequent task that has been extensively studied in decision analysis: the selection of a subset of preferred objects (e.g., houses) out of a set of possible alternatives. In our evaluation framework, the user performs this task by using a system for interactive data exploration and analysis (IDEA), see Figure 3. Let's now examine how GEA can be evaluated in the context of the selection task, by going through the evaluation framework architecture.

3.1 The Evaluation Framework Architecture

As shown in Figure 2, the evaluation framework 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. At the onset, the user is assigned the task to select from the dataset the four most preferred alternatives and to place them in a Hot List (see Figure 3, upper right corner) ordered by preference. Whenever the user feels that the task is accomplished, the ordered list of preferred alternatives is saved as her Preliminary Hot List (Figure 2 (2)). After that, this list and the initial Model of User's Preferences are analysed by the User Model Refiner to produce a Refined Model of the User's Preferences (Figure 2 (3)). 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 (Figure 2 (4)). 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 (Figure 2 (5)), which is presented to the user by the IDEA system (Figure 2 (6))(see also Figure 3 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 immediately to introduce NewI in her Hot List, or (b) decide to further explore the dataset, possibly making changes and adding NewI to the Hot List, or (c) do nothing.

Figure 3 shows the display at the end of the interaction, when the user, after reading the argument, has decided to introduce NewI in the Hot List first position (Figure 3, top right).

Whenever the user decides to stop exploring and is satisfied with her final selection, measures related to argument's effectiveness can be assessed (Figure 2 (7)). These measures are obtained either from the record of the user interaction with the system or from user self-reports in a final questionnaire (see Figure 4 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 and (c) how much she likes NewI and the other objects in the Hot List.

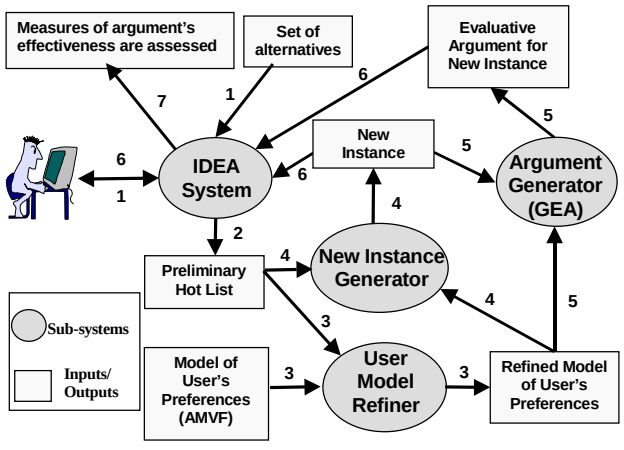

Figure 2 The evaluation framework architecture

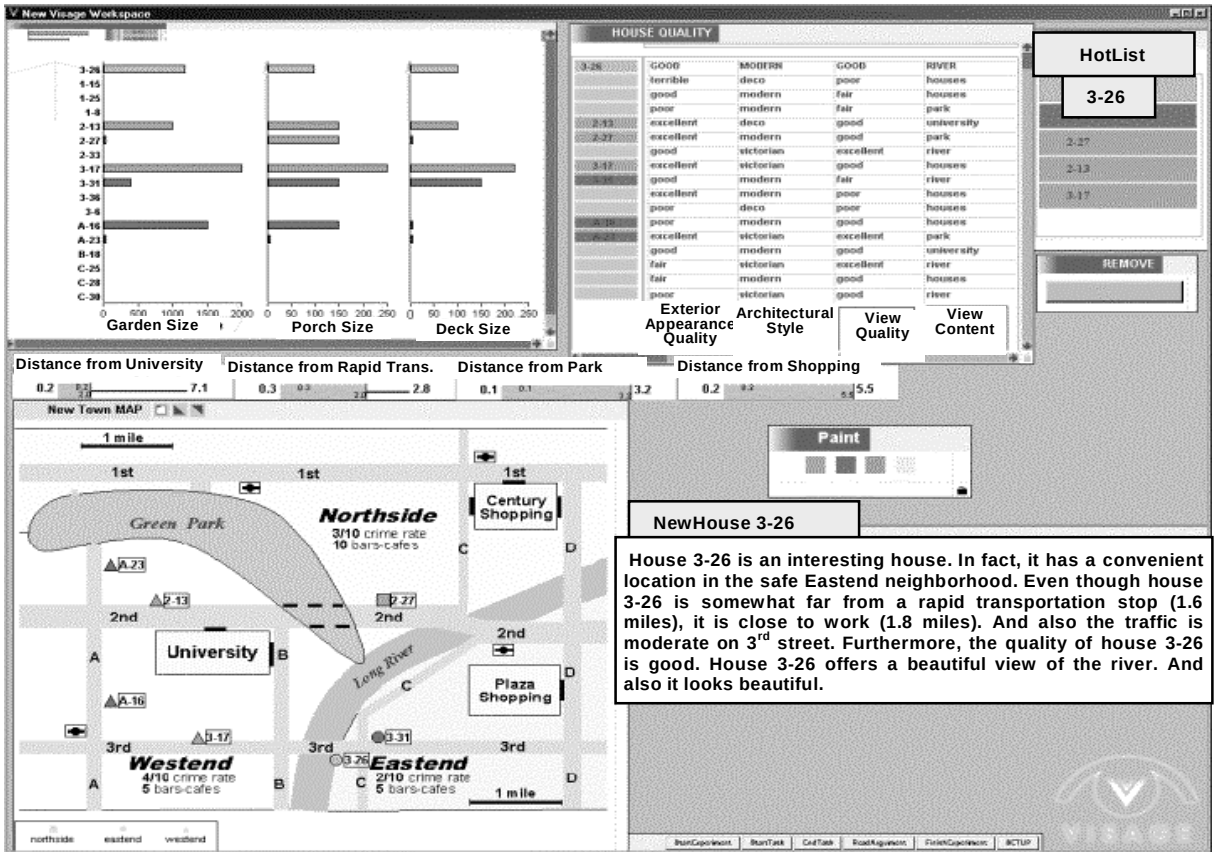

Figure 3 The IDEA environment display at the end of the interaction

- A measure of the user's confidence that she has selected the best for her in the set of alternatives.
- A measure of argument effectiveness derived by explicitly questioning the user at the end of the interaction about the rationale for her decision [Olso and Zanna 1991]. This can provide valuable information on what aspects of the argument were more influential on the user's decision.
- An additional measure of argument effectiveness is 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 behavioural and attitudinal changes [Olso and Zanna 1991].

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 and measurements are collected on its effectiveness. We now discuss an experiment we have performed within the evaluation framework to test to what extent tailoring an evaluative argument to a model of the user preferences increases its effectiveness.

4 The Experiment

Given the goal of our empirical investigation, we have performed a between-subjects experiment with three experimental conditions: (i) *No-Argument* - subjects are simply informed that NewI came on the market. (ii) *Tailored* - subjects are presented with an evaluation of NewI tailored to their preferences. (iii) *Non-Tailored* - subjects are presented with an evaluation of NewI 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 [Srivastava, Connolly et al. 1995]. In the three conditions, all the information about the NewI is also presented graphically, so that no information is hidden from the subject.

Our hypotheses on the experiment are the following. First, we expect arguments generated for the Tailored condition to be more effective than arguments generated for the Non-Tailored condition. Second, the Tailored condition should be somewhat better than the No-Argument condition, but to a lesser extent, because subjects, in the absence of any argument, may spend more time further exploring the

dataset, thus reaching a more informed and balanced decision. Finally, we do not have strong hypotheses on comparisons of argument effectiveness between the No-Argument and Non-Tailored conditions. The experiment is organized in two phases. In the first phase, the subject fills out a questionnaire on the Web which implements a method from decision theory to acquire an AMVF model of the subject's preferences [Edwards and Barron 1994]. In the second phase, to control for possible confounding variables, including subject's argumentativeness [Infante and Rancer 1982], need for cognition [Cacioppo, Petty et al. 1983], intelligence and self-esteem, the subject is randomly assigned to one of the three conditions. Then, the subject interacts with the evaluation framework and at the end of the interaction measures of the argument effectiveness are collected, as described in Section 3.1.

After running the experiment with 8 pilot subjects to refine and improve the experimental procedure, we ran a formal experiment involving 30 subjects, 10 in each experimental condition. Each subject had only one interactive session with the framework.

5 Experiment Results

5.1 A Precise Measure of Satisfaction

According to literature on persuasion, the most important measures of argument effectiveness are the ones of behavioral intentions and attitude change [Olso and Zanna 1991]. As explained in Section 3.1, in our framework these measures include (a) whether or not the user adopts NewI, (b) in which position in the Hot List she places it, (c) how much she likes the proposed 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) How would you judge the houses in your Hot List?
The more you like the house the closer you should put a cross to *"good choice"*

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*

Figure 4 Sample filled-out self-report on user's satisfaction with houses in the Hot List[1]

[1] If the subject does not adopt the new house, she is asked to express her satisfaction with the new house in an additional self-report.

A closer analysis of the above measures 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 Figure 4, the subject was able to specify that the first house in the Hot List was only one space (unit of satisfaction) better then the house following it in the ranking, while the third house was two spaces better than the house following it.

(ii) Self-reports do not force subjects to express a total order between the houses. For instance, in Figure 4 the subject was allowed to express that the second and the third house 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. A z-score is a normalized distance in standard deviation units of a measure x_i from the mean of a population X. Formally:

$x_i \in X$; $z\text{-}score(x_i, X) = [x_i - \mu(X)] / \sigma(X)$

For instance, the satisfaction z-score for the new instance, given the sample self-reports shown in Figure 4, would be:
$[7 - \mu(\{8,7,7,5\})] / \sigma(\{8,7,7,5\}) = 0.2$
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.

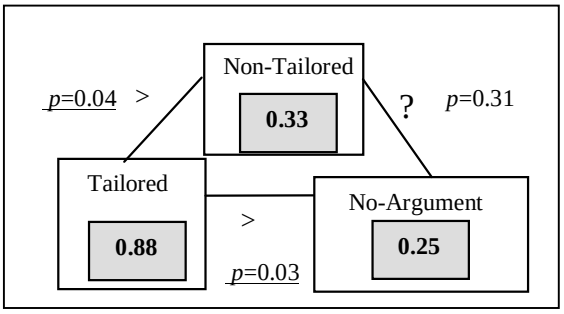

Figure 5 Results for satisfaction z-scores. The average z-scores for the three conditions are shown in the grey boxes

5.2 Results

As shown in Figure 5, the satisfaction z-scores obtained in the experiment confirmed our hypotheses. Arguments generated for the Tailored condition were significantly more effective than arguments generated for the Non-Tailored condition (p=0.04). The Tailored condition was also

significantly better than the No-Argument condition (p=0.03). And this happened despite the fact that subjects in the No-Argument condition spent significantly more time further exploring the dataset after NewI was presented (as indicated in Table 1)(p=0.05). Finally, we found no significant difference in argument effectiveness between the No-Argument and Tailored-Verbose conditions.

With respect to the other measures of argument effectiveness mentioned in Section 3.1, we have not found any significant differences among the experimental conditions[2].

No-Argument	Non-Tailored	Tailored
0:03:56	0:03:37	0:02:44

Table 1 Average time spent by subjects in the three conditions further exploring the dataset after the new house is presented (from Logs of the interaction).

6 Conclusions and Future Work

Argumentation theory indicates that effective arguments should be tailored to a model of the user's preferences. Although previous work on automatically generating evaluative arguments has followed this basic indication, the effect of tailoring on argument effectiveness has never been measured empirically. As an initial attempt to address this issue, we have compared in a formal experiment the effectiveness of arguments that were tailored vs. non-tailored to a model of the user preferences. The experiment results show that tailored arguments are significantly better.

As future work, we plan to perform further experiments. We intend to repeat the same experiment in a different domain to test for external validity. We also envision an experiment in conditions of uncertainty, in which we may compare arguments tailored to an AMVF, with argument tailored to more sophisticated models of user's preferences, that consider interactions among objectives.

References

[Cacioppo, Petty et al. 1983] Cacioppo, J. T., R. E. Petty, et al. Effects of Need for Cognition on Message Evaluation, Recall, and Persuasion. *Journal of Personality and Social Psychology* 45(4): 805-818, 1983.

[Carenini 2000] Carenini, G. Generating and Evaluating Evaluative Arguments. Ph.D. Thesis, Intelligent System Program, University of Pittsburgh, 2000.

[Carenini 2000] Carenini, G. A Task-based Framework to Evaluate Evaluative Arguments. *First International Conference on Natural Language Generation*, Mitzpe Ramon, Israel: 9-16, 2000.

[Carenini and Moore 2000] Carenini, G. and J. Moore. A Strategy for Generating Evaluative Arguments. *First International Conference on Natural Language Generation*, Mitzpe Ramon, Israel: 47-54, 2000.

[Chai, Budzikovaska et al. 2000] Chai, J., M. Budzikovaska, et al. Natural Language Sales Assistant. *38th Annual Meeting of the Association for Computational Linguistics*: 34-35, 2000.

[Clemen 1996] Clemen, R. T. *Making Hard Decisions: an introduction to decision analysis*. Belmont, California, Duxbury Press, 1996.

[Edwards and Barron 1994] Edwards, W. and F. H. Barron. SMARTS and SMARTER: Improved Simple Methods for Multiattribute Utility Measurements. *Organizational Behavior and Human Decision Processes* 60: 306-325, 1994.

[Elhadad, McKeown et al. 1997] Elhadad, M., K. McKeown, et al. Floating constraints in lexical choice. *Computational Linguistics* 23(2): 195-239, 1997.

[Elzer, Chu-Carroll et al. 1994] Elzer, S., J. Chu-Carroll, et al. Recognizing and Utilizing User Preferences in Collaborative Consultation Dialogues. *Proc. of Fourth Int. Conf. of User Modeling*, Hyannis, MA: 19-24, 1994.

[Infante and Rancer 1982] Infante, D. A. and A. S. Rancer. A Conceptualization and Measure of Argumentativeness. *Journal of Personality Assessment* 46: 72-80, 1982.

[Jameson, Schafer et al. 1995] Jameson, A., R. Schafer, et al. Adaptive provision of Evaluation-Oriented Information: Tasks and techniques. *Proceedings of 14th IJCAI*, Montreal, 1995.

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

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

[Morik 1989] Morik, K. *User Models and Conversational Settings: Modeling the User's Wants*. User Models in Dialog Systems. A. Kobsa and W. Wahlster, Springer-Verlag**:** 364-385, 1989.

[Olso and Zanna 1991] Olso, J. M. and M. P. Zanna. *Attitudes and beliefs: Attitude change and attitude-behavior consistency*. Social Psychology. R. M. Baron and W. G. Graziano, 1991.

[Solomon 1998] Solomon, M. R. *Consumer Behavior: Buying, Having, and Being*, Prentice Hall, 1998.

[Srivastava, Connolly et al. 1995] Srivastava, J., T. Connolly, et al. Do Ranks Suffice? A Comparison of Alternative Weighting Approaches in Value Elicitation. *Organizational Behavior and Human Decision Process* 63(1): 112-116, 1995.

[2] The measure of decision rationale is still under analysis.