Evaluating User-Tailored Evaluative Arguments: Persuasive Effectiveness and its Communicative Appropriateness

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

Abstract

Computer-based persuasive technologies could soon become powerful tools of social influence. It is therefore critical that they are thoroughly tested before being adopted in practice. However, evaluating these technologies requires being clear about what effective persuasion actually means. In this paper, we argue that effective persuasion can have a different meaning depending on the communicative context. In light of this consideration, we discuss further analysis of the results of an empirical evaluation of GEA, a computer system that generates evaluative arguments tailored to a quantitative model of the user preferences.

1 Introduction

Persuasion is an attempt to reinforce or change someone's attitudes and behaviors about an issue, object or action. With the ever-increasing availability of computer systems, their study as persuasive technologies has recently generated considerable interest [Fogg 1999] [Fogg 2003]. This new field has been even given a brand new name: captology, which is an acronym based on the phrase "computers as persuasive technologies". It is clear that, if used according to sound ethical principles, widespread persuasive computer systems could generate tremendous benefits in many domains, including health and education, by changing their users' attitudes (and hopefully their behavior) in the desired directions. However, persuasive systems could also be used unethically. They could be used, for instance, to persuade people to buy products they do not really like or need [Greer, MacKenzie et al. 1996]. Because of the potential power of persuasive technologies as tools of social influence, and the ethical implications of their adoption, it is crucial that persuasive technologies are thoroughly analyzed and tested before being adopted in practice.

Research in user modeling plays a key role in developing persuasive technologies because, as persuasion theory indicates [Miller and Levine 1996], personalization (i.e., the tailoring of information to the intended audience), is a very effective persuasive strategy. Even if we limit our attention to persuasive evaluative presentations about a single entity (i.e, arguments claiming that the entity is good vs. bad for the user), several systems have been presented in the literature that generate personalized arguments tailored to a model of the user preferences [Elhadad 1995], [Ardissono and Goy 2001], [Morik 1989], [Elzer, Chu-Carroll et al. 1994], [Carenini and Moore 2000b].

Unfortunately, despite the importance of empirical testing for persuasive technology, most of the research in user modeling that can be applied to persuasion has not been formally evaluated. Furthermore, when empirical testing has been performed, it has considered a rather limited notion of persuasive effectiveness. In particular, the effectiveness of an evaluative argument about an entity has been measured as the degree to which the attitude of the user toward the entity is moved in the direction supported by the argument (e.g., [Carenini and Moore 2001]).

In this paper we propose a revision of this notion of persuasive effectiveness by considering that effective persuasion can mean different things in different communicative contexts. For instance, in some situations the more the user is persuaded of something the better, while in other situations only a given degree of persuasion is desirable. As an example of the first case, consider an argument attempting to persuade the user to follow a healthy diet. Here, the user wants to follow an healthy diet in spite of her food preferences, so he is willing to be persuaded as much as possible. In contrast, an example of the second type could be an argument attempting to convince the user that a given financial portfolio is the best choice for her. Here, the user would probably like to be persuaded only to a degree consistent with her preferences and risk-attitude.

In the remainder of the paper, we first present few examples of persuasive communication that further clarify why different measures of persuasive effectiveness are appropriate in different communicative context. Then we describe our generator of evaluative arguments (GEA): a system that automatically generates arguments tailored to a quantitative model of the user preferences. After that, we summarize our initial empirical testing of GEA [Carenini and Moore 2000a; Carenini and Moore 2001], in which we assumed a restrictive definition of argument effectiveness. Finally, we present a revision of that evaluation in light of our novel and more comprehensive notion of persuasive effectiveness.

2 Persuasive Effectiveness in Different Communicative Contexts

In this paper we argue that the effectiveness of an evaluative argument should be measured differently depending on the nature of the communicative context. For illustration, we now examine four prototypical communicative contexts, which differ with respect to how the effectiveness of an argument generated in that context should be measured.

Advertisement – the goal of an advertisement is to convince the audience to buy or patronize something (out of all possible alternatives). This is a completely persuasive goal, thus, the absolute degree of positive increase of the audience attitude towards the proposal is an appropriate measure of effectiveness.

Salesperson helping a user to select a product - The salesperson is collaborating with the customer because she wants him to buy a product that he is going to like and enjoy. But, at the same time, she has the goal of increasing as much as possible the monetary value of his purchase (by stretching the customer's initial budget). And maybe she also has the goal of pushing some lines of products (e.g., to balance the inventory). Thus, whenever a salesman generates an evaluative argument about a product, she is aiming at increasing the customer's positive attitude towards the product, but not excessively so, because that might generate false expectations and possibly, later on, customer frustration.

Doctor and patient revising treatment selection- One of the most communication intensive decisions in medicine is the treatment choice that follows diagnosis. Typically, the doctor helps the patient to choose a treatment out of the available ones, by taking into account the patient's preferences. Obviously, in the selection process any evaluative arguments generated by the doctor should try to move the user's attitude towards the treatments as close as possible to the level indicated by the patient's preferences. However, once the selection has been made, a process of revision of the selected treatment might be more persuasive in nature. In fact, in revising the decision with the patient, the doctor may generate an evaluative argument for the selected treatment that may aim at slightly boosting the patient's attitude towards that treatment. The rationale being that by increasing the patient's positive attitude towards the treatment, the doctor may increase patient's compliance, and consequently the effectiveness of the treatment.

Pure Advisor in any domain - A pure advisor has no other goal except for helping the user to select the best entity for her out of a set of available alternatives. When the user asks for an evaluation of an entity, the advisor should generate an argument that moves the user's attitude towards the entity as close as possible to the level indicated by the user's preferences.

3 Generating Evaluative Arguments

Our generation system, known as the Generator of Evaluative Arguments (GEA), 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 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 pipelined generation system, including a text planner [Young and Moore 1994], a microplanner (which integrates and extends previous work [Elhadad 1991; Grosz, Joshi et al. 1995; Shaw 1998] and a sentence realizer [Elhadad and Robin 1996].

3.1 Background on AMVF and its Usage 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 [Carenini and Moore 2000b]. First, GEA can compute how valuable any objective of the

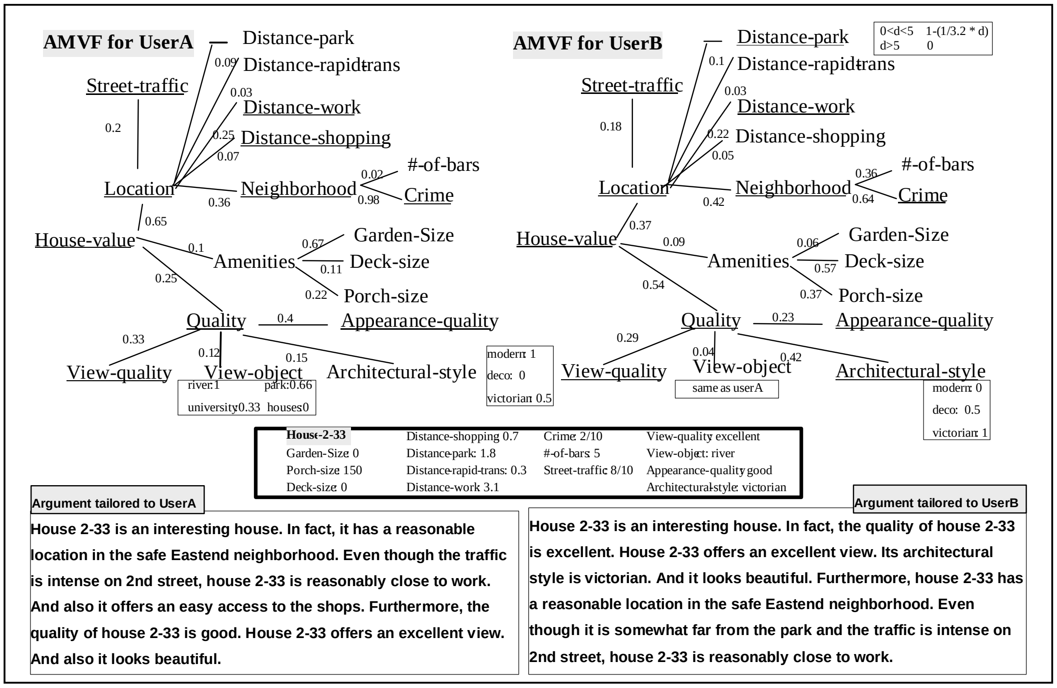

Figure 1 Top: AMVF for two sample users; for clarity's sake only a few component value functions are shown. Bottom: arguments about house-2-33, tailored to the two different models

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 toward the evaluation of its parent objective. In this way, in compliance with argumentation theory, evidence can be arranged according to its strength. Furthermore, arguments of different length can be generated in a principled way by including only evidence whose strength is above a given threshold [Carenini and Moore 2000a]. 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 AMVFs 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.

3.2 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 AMVF models representing 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 box 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* of the target house is *good* for UserA, but *excellent* for UserB.

4 GEA Empirical Evaluation

The evaluation of GEA was based on the *task efficacy* evaluation method. This method allows the experimenter to evaluate a generation system 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, 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.

4.1 The Evaluation Framework

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 (Figure 2 (1)), the user is assigned the task to select from a 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

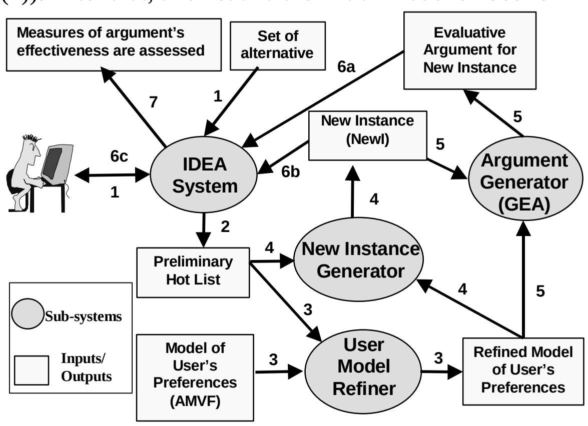

Figure 2 The evaluation framework architecture

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 (6a)). Figure 3 (bottom-right) shows an argument generated for the NewI house-3-26. The argument goal is to persuade the user that NewI is worth being considered.

All the information about NewI is also presented graphically (Figure 2 (6b)). 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 (Figure 2 (6c)), 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.
- 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 convincigness. 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: (i) to what extent tailoring an evaluative argument to a model of the user preferences increases its effectiveness; (ii) the influence of argument conciseness on argument effectiveness.

4.2 The Experiment

Given the goal of our empirical investigation, we performed a between-subjects experiment with four experimental conditions: (i) *No-Argument* - subjects are simply informed that NewI came on the market. (ii) *Tailored-Concise*: subjects are presented with an evaluation of NewI tailored to their preferences and at a level of conciseness that we hypothesize to be optimal. (iii) *Tailored-Verbose*: subjects are presented with an evaluation of NewI tailored to their preferences, but at a level of conciseness that we hypothesize to be too low. (iv) *Non-Tailored-Concise* - 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). The level of conciseness in this condition is the one that we hypothesize to be optimal.

Notice that in the four conditions, all the information about the NewI is also presented graphically (see Figure 3), 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-Concise condition to be more effective than arguments generated for the Non-Tailored-Concise and Tailored-Verbose conditions. Second, the Tailored-Concise

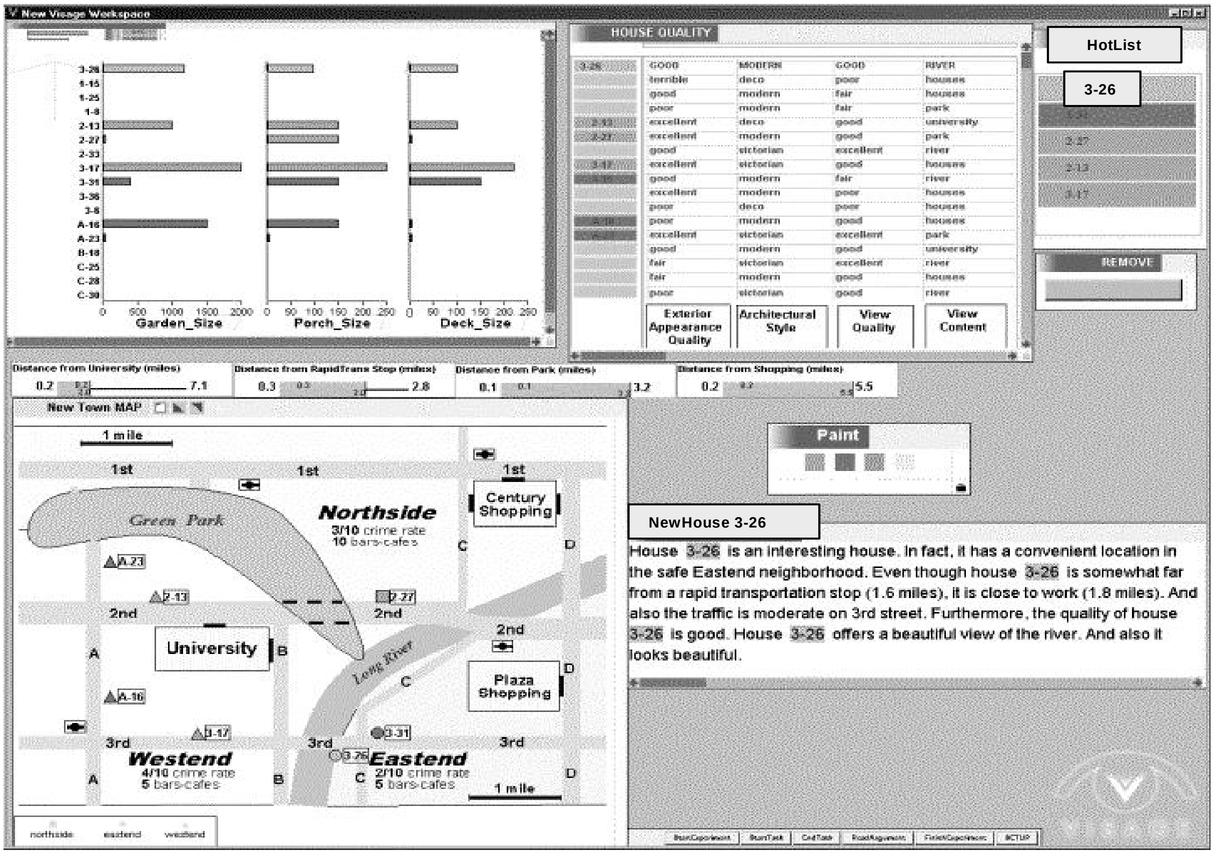

Figure 3 The IDEA environment display at the end of the interaction

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, Non-Tailored-Concise and Tailored-Verbose 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, the subject is randomly assigned to one of the three conditions in order 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. 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 4.1.

After running the experiment with 8 pilot subjects to refine and improve the experimental procedure, we ran a formal experiment involving 40 subjects, 10 in each experimental condition.

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 4.1, 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 (Figure 4).

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.

Section 4.1, we have not found any significant differences among the experimental conditions.

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 they assess, like (a) and (b), a preference ranking among the new instance 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 preceding it in the ranking, while the third house was two spaces better than the house preceding 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.

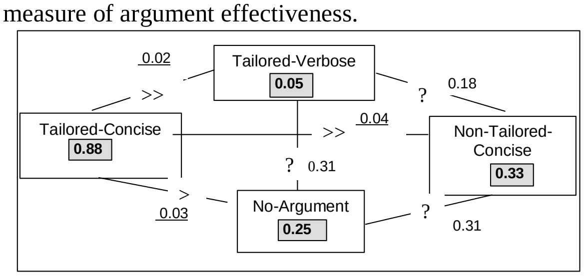

Figure 5 Results for satisfaction z-scores. The average z-scores for the four conditions are shown in the grey boxes. T-test levels are shown on the links (underlined when significant)

As shown in Figure 5, the satisfaction z-scores obtained in the experiment confirmed our main hypotheses. Arguments generated for the Tailored-Concise condition were significantly more effective than arguments generated both for the Non-Tailored-Concise condition (p=0.04) and for the Tailored-Verbose condition (p=0.02). The Tailored condition was also significantly better than the No-Argument condition (p=0.03). With respect to the other measures of argument effectiveness mentioned in Section 4.1, we have not found any significant differences among the experimental conditions.

5 A Revision of GEA Empirical Evaluation

In the evaluation of GEA, reported in the previous section, we considered a rather limited notion of persuasive effectiveness. Essentially, the effectiveness of an evaluative argument about an entity was measured (by means of the satisfaction z-scores) as the degree to which the attitude of the user toward the entity is moved in the direction supported by the argument. However, as we have previously discussed in Section 2, although this measure is reasonable in certain communicative contexts (e.g., advertisement), it becomes less and less appropriate the more the communicative context moves from persuading to advising. In fact, in an advising situation the goal of an evaluative argument should not be to increase the user's attitude towards the argument subject as much as possible, but only to a level that is consistent with the user's preferences. In essence, a pure advisor should not try to convince you that you like something more than your preferences would indicate.
In light of this observation, we have revised the results of the GEA evaluation to verify how arguments in the four experimental conditions fare with respect to the measure of effectiveness that is more appropriate in advising situations. To address this question, we must compare the user's attitude towards the new instance once the argument is presented, with how valuable the new instance is for the user according to the model of her preferences (represented in GEA as and AMVF – see Section 3.1).
Table 1 presents the results of this analysis. In the four conditions, the satisfaction z-score of the new instance (obtained through self-report) is compared with the z-score of the new instance value according to the user's preference model[2].

Condition	Average z-score new instance value according to model	Average z-score new instance satisfaction (self-report)	Difference	T-test p-value
No-Argument	0.51	0.25	-0.26	0.16
Non-Tailored-Concise	0.56	0.28	-0.22	0.14
Tailored-Concise	0.64	0.88	+0.24	0.14
Tailored-Verbose	0.54	0.05	-0.49	0.08

Table 1 Differences between expected value and satisfaction in the four experimental conditions

Unfortunately, in this analysis, we only obtained one weakly significant result. As shown in the table,

[2] Both z-scores are computed with respect to the instances in the Hot-List

arguments generated in the Tailored-Verbose condition seem to be inappropriate in a pure advising situation because they move the user attitude towards the subject below the level indicated by the user's preferences (-.49). For the other three conditions, the differences are not significant, therefore any interpretations must be taken as even more tentative and preliminary. From the trends that we can observe in the table, it appears that both No-Argument and Non-Tailored-Concise conditions would be inappropriate in an advising situation. Similarly to the Tailored-Verbose condition, they do not increase enough the user attitude towards the subject (-.26 and -.22 respectively). In contrast, Tailored-Concise seems to be inappropriate for the opposite reason. It pushes the user attitude too high (+.24).

In light of these results, it remains an open issue what condition would generate the most effective arguments in an advising situation. Since the decrement in argument conciseness from the Tailored-Concise to the Tailored-Verbose condition reduced the user satisfaction z-score (second column) from 0.88 to 0.05 and consequently the difference between value and satisfaction (third column) from +0.24 to –0.49, it seems reasonable that by decreasing the conciseness of arguments in the Tailored-Concise condition (by a smaller decrement than in Tailored-Verbose), we might reduce the difference (from +0.24 to 0). Obviously, further experiments are needed to verify this hypothesis.

6 Conclusions and Future Work

Computer systems that generate evaluative arguments tailored to a model of the user's preferences should be thoroughly tested before being adopted, because of their potential power as tools of social influence.

In this paper, we have argued that the persuasive effectiveness of user tailored evaluative arguments should not be measured only in absolute terms, i.e., as the degree to which the attitude of the user toward the entity is moved in the direction supported by the argument. Rather, the user's attitude change toward the entity, generated by the argument, should be compared (whenever possible) with the attitude toward the entity predicted by the model of the user's preferences. The reason is that, in different communicative contexts, arguments that generate an attitude equal-to, less-than, or greater-than the one predicted by the model are more or less desirable.

The revision of the analysis of GEA evaluation outcomes we have discussed in this paper, is just an example of how this kind of more informative evaluations should be performed. In general, whenever the expected persuasive force of an argument generator system can be varied by changing a set of parameters of the system (e.g., the threshold on the measure of evidence strength that defines when a piece of evidence is worth mentioning in GEA), the following methodology should be applied. First, experiments should be run to verify to what extent the attitude of the user toward an entity is increased by arguments generated for different settings of the system's parameters. Then, by comparing these measurements with the user attitude towards the entity predicted by the user model, it should be possible to determine what system parameter settings are more or less appropriate in different communicative contexts. For instance, if an argument generation system with parameter setting *ps* generates arguments that increase the user attitude toward an entity to a level much greater than the one predicted by the user model, that system with setting *ps* should be very effective in an advertisement situation.

We have applied this methodology in our new analysis of GEA evaluation outcomes. Although we have not found any statistically significant difference in the data collected, we plan to run further studies to verify the trends we have identified. For instance, simply repeating the experiment we have described with more subjects might generate significant results.

We also plan to apply the methodology to test other aspects of GEA. For instance, we may run experiments to test the effectiveness of alternative argumentation strategies (different from the one we used in our evaluation and described in [Carenini and Moore 2000a]). An argumentation strategy specifies what content should be included in an argument and how the selected content should be ordered. So, we expect different argumentation strategies to be more or less appropriate in different contexts.

Finally, the discussion (in Section 2) of the classification of communicative contexts and their corresponding measures of persuasive effectiveness was rather informal and based essentially on our intuitions. The reason for this was that we could not find any detailed and formal treatment of the issue in argumentation literature. As a next step of our research, we intend to further investigate our preliminary classification of communicative contexts and the notion that effective persuasion is context specific. We hope to obtain useful suggestions and pointers to relevant literature, as we will present our research plan at the workshop.

References

[Ardissono and Goy 2001] L. Ardissono and A. Goy. Tailoring the Interaction with Users in Web Stores. *User Modeling and User-Adapted Interaction (UMUAI)*, 2001.

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

[Carenini and Moore 2000a] G. Carenini and J. Moore. An Empirical Study of the Influence of Argument Conciseness on Argument Effectiveness. Proc. of the *Annual Meeting of the Association for Computational Linguistics (ACL)*, Hong Kong, China: 150-157, 2000.

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

[Carenini and Moore 2001] G. Carenini and J. Moore. An Empirical Study of the Influence of User Tailoring on Evaluative Argument Effectiveness. Proc. of the *17th*

International Joint Conference on Artificial Intelligence, Seattle, USA, 2001.

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

[Edwards and Barron 1994] W. Edwards 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 1991] M. Elhadad. *Generating Adjectives to Express the Speaker's Argumentative Intent*. Proceedings of the 9th National Conference on Artificial Intelligence. Anaheim**:** 98--104, 1991.

[Elhadad 1995] M. Elhadad. Using argumentation in text generation. *Journal of Pragmatics* 24: 189-220, 1995.

[Elhadad and Robin 1996] M. Elhadad and J. Robin. An overview of SURGE: a reusable comprehensive syntactic realization component, Dept of Mathematics and Computer Science, Ben Gurion University, Beer Sheva, Israel, 1996.

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

[Fogg 1999] B. J. Fogg. Persuasive Technologies. *Communications of the ACM* 42(5): 27-29, 1999.

[Fogg 2003] B. J. Fogg. *Persuasive Technology: Using Computers to Change What We Think and Do*, Morgan Kaufmann Publishers, 2003.

[Greer, MacKenzie et al. 1996] J. E. Greer, M. MacKenzie and G. M. Koehn. User Models for Persuasion, Coercion and Sales, ARIES Laboratories, Dept. of Computer Science, University of Saskathewan, 1996.

[Grosz, Joshi et al. 1995] B. J. Grosz, A. K. Joshi and S. Weinstein. Centering: A Framework for Modelling the Local Coherence of Discourse. *Computational Linguistics* 21(2): 203-226, 1995.

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

[Jameson, Schafer et al. 1995] A. Jameson, R. Schafer, J. Simons and T. Weis. Adaptive provision of Evaluation-Oriented Information: Tasks and techniques. Proc. of the *Proceedings of 14th IJCAI*, Montreal: 1886-1893, 1995.

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

[Miller and Levine 1996] M. D. Miller and T. R. Levine. *Persuasion*. An Integrated Approach to Communication Theory and Research. M. B. Salwen and D. W. Stack. Mahwah, New Jersey**:** 261-276, 1996.

[Morik 1989] K. Morik. *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] J. M. Olso and M. P. Zanna. *Attitudes and beliefs: Attitude change and attitude-behavior consistency*. Social Psychology. R. M. Baron and W. G. Graziano, 1991.

[Shaw 1998] J. Shaw. Clause Aggregation Using Linguistic Knowledge. Proc. of the *9th International Workshop on Natural Language Generation*, Niagara-on-the-Lake, Canada: 138-147, 1998.

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

[Young and Moore 1994] R. M. Young and J. D. Moore. DPOCL: A Principled Approach to Discourse Planning. Proc. of the *Proceedings of the 7th International Workshop on Text Generation*, Montreal, Canada, 1994.