A Task-based Framework to Evaluate Evaluative Arguments

Giuseppe Carenini
Intelligent Systems Program
University of Pittsburgh, Pittsburgh, PA 15260, USA
carenini@cs.pitt.edu

Abstract

We present an evaluation framework in which the effectiveness of evaluative arguments can be measured with real users. The framework is based on the task-efficacy evaluation method. An evaluative argument is presented in the context of a decision task and measures related to its effectiveness are assessed. Within this framework, we are currently running a formal experiment to verify whether argument effectiveness can be increased by tailoring the argument to the user and by varying the degree of argument conciseness.

Introduction

Empirical methods are fundamental in any scientific endeavour to assess progress and to stimulate new research questions. As the field of NLG matures, we are witnessing a growing interest in studying empirical methods to evaluate computational models of discourse generation (Dale, Eugenio et al. 1998). However, with the exception of (Chu-Carroll and Carberry 1998), little attention as been paid to the evaluation of systems generating evaluative arguments, communicative acts that attempt to affect the addressee's attitudes (i.e. evaluative tendencies typically phrased in terms of like and dislike or favor and disfavor).

The ability to generate evaluative arguments is critical in an increasing number of online systems that serve as personal assistants, advisors, or sales assistants[1]. For instance, a travel assistant may need to compare two vacation packages and argue that its current user should like one more than the other.

In this paper, we present an evaluation framework in which the effectiveness of evaluative arguments can be measured with real users. The measures of argument effectiveness used in our framework are based on principles developed in social psychology to study persuasion (Miller and Levine 1996). We are currently applying the framework to evaluate arguments generated by an argument generator we have developed (Carenini 2000). To facilitate the evaluation of specific aspects of the generation process, the argument generator has been designed so that its functional components can be easily turned-off or changed.

In the remainder of the paper, we first describe our argument generator. Then, we summarize literature on persuasion from social psychology. Next, we discuss previous work on evaluating NLG models. Finally, we describe our evaluation framework and the design of an experiment we are currently running.

1 The Argument Generator

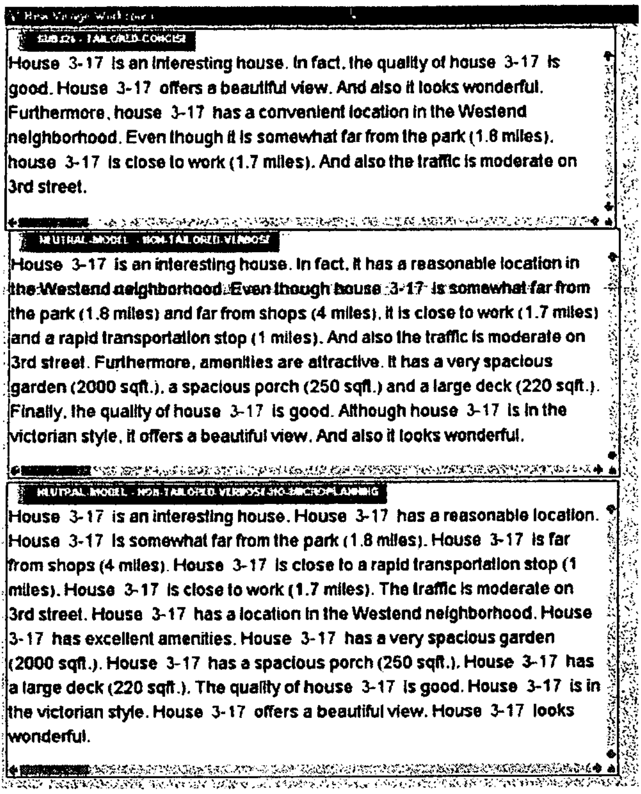
The architecture of the argument generator is a typical pipelined architecture comprising a discourse planner, a microplanner and a sentence realizer.

The input to the planning process is an abstract evaluative communicative action expressing:

- The subject of the evaluation, which can be an entity or a comparison between two entities in the domain of interest (e.g., a house or a comparison between two houses in the real-estate domain).
- An evaluation, which is a number in the interval [0,1] where, depending on the subject, 0 means "terrible" or "much worse" and 1 means "excellent" or "much better").

Given an abstract communicative action, the *discourse planner* (Young and Moore 1994) selects and arranges the content of the argument

[1] See for instance www.activebuyersguide.com


SUBJ26 - TAILORED-CONCISE
NEUTRAL-MODEL - NON-TAILORED-VERBOSE
NEUTRAL-MODEL - NON-TAILORED-VERBOSE-NO-MICROPLANNING

House 3-17 is an interesting house. In fact, the quality of house 3-17 is good. House 3-17 offers a beautiful view. And also it looks wonderful. Furthermore, house 3-17 has a convenient location in the Westend neighborhood. Even though it is somewhat far from the park (1.8 miles), house 3-17 is close to work (1.7 miles). And also the traffic is moderate on 3rd street.

House 3-17 is an interesting house. In fact, it has a reasonable location in the Westend neighborhood. Even though house 3-17 is somewhat far from the park (1.8 miles) and far from shops (4 miles), it is close to work (1.7 miles) and a rapid transportation stop (1 miles). And also the traffic is moderate on 3rd street. Furthermore, amenities are attractive. It has a very spacious garden (2000 sqft.), a spacious porch (250 sqft.) and a large deck (220 sqft.). Finally, the quality of house 3-17 is good. Although house 3-17 is in the victorian style, it offers a beautiful view. And also it looks wonderful.

House 3-17 is an interesting house. House 3-17 has a reasonable location. House 3-17 is somewhat far from the park (1.8 miles). House 3-17 is far from shops (4 miles). House 3-17 is close to a rapid transportation stop (1 miles). House 3-17 is close to work (1.7 miles). The traffic is moderate on 3rd street. House 3-17 has a location in the Westend neighborhood. House 3-17 has excellent amenities. House 3-17 has a very spacious garden (2000 sqft.). House 3-17 has a spacious porch (250 sqft.). House 3-17 has a large deck (220 sqft.). The quality of house 3-17 is good. House 3-17 is in the victorian style. House 3-17 offers a beautiful view. House 3-17 looks wonderful.

Figure 1 Sample arguments in order of decreasing expected effectiveness for the target user SUBJ-26

by decomposing abstract communicative actions into primitive ones and by imposing appropriate ordering constraints among communicative actions. Two knowledge sources are involved in this process:

- A complex model of the user's preferences based on multiattribute utilility theory (MAUT)(Clemen 1996).
- A set of plan operators, implementing guidelines for content selection and organisation from argumentation theory (Carenini and Moore 2000).

By using these two knowledge sources, the discourse planner produces a text plan for an argument whose content and organization are tailored to the user according to argumentation theory.

Next, the text plan is passed to the *microplanner* which performs aggregation, pronominalization and makes decisions about cue phrases. Aggregation is performed according to heuristics similar to the ones proposed in (Shaw 1998). For pronominalization, simple rules based on centering are applied (Grosz, Joshi et al. 1995). Finally, decisions about cue phrases are made according to a decision tree based on suggestions from (Knott 1996; di Eugenio, Moore et al. 1997) . The *sentence realizer* extends previous work on realizing evaluative statements (Elhadad 1995).

The argument generator has been designed to facilitate the testing of the effectiveness of different aspects of the generation process. The experimenter can easily vary the expected effectiveness of the generated arguments by controlling whether the generator tailors the argument to the current user, the degree of conciseness of the generated arguments and what microplanning tasks are performed.

Figure 1 shows three arguments generated by the argument generator that clearly illustrate this feature. We expect the first argument to be very effective for the target user. Its content and organization has been tailored to her preferences. Also, the argument is reasonably fluent because of aggregation, pronominalization and cue phrases. In contrast, we expect the second argument to be less effective with our target user, because it is not tailored to her preferences[2], and it appears to be somewhat too verbose[3]. Finally, we expect the third arguments not to be effective at all. It suffers from all the shortcomings of the second argument, with the additional weakness of not being fluent (no microplannig tasks were performed).

2 Research in Psychology on Persuasion

Arguing an evaluation involves an intentional communicative act that attempts to affect the current or future behavior of the addressees by creating, changing or reinforcing the addressees' attitudes. It follows that the effectiveness of an evaluative argument can be tested by comparing measurements of subjects' attitudes or behavior before and after their exposure to the argument. In many experimental situations, however, measuring effects on overt behavior can be problematic (Miller and Levine 1996), therefore most research on persuasion has been based either on measurements of attitudes or on declaration of behavioral intentions. The most common technique to measure attitudes is subject self-report (Miller and Levine 1996). Typically, self-report measurements involve the use of a scale that consists of two ``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 Figure 4 for examples).

Research in persuasion suggests that some individuals may be naturally more resistant to persuasion than others (Miller and Levine 1996). Individual features that seem to matter are: *argumentativeness* (tendency to argue)(Infante and Rancer 1982), *intelligence*, *self-esteem* and *need for cognition* (tendency to engage in and to enjoy effortful cognitive endeavours)(Cacioppo, Petty et al. 1983). Any experiment in persuasion should control for these variables.

A final note on the evaluation of arguments. An argument can also be evaluated by the argument addressee with respect to several 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 changes (Olso and Zanna 1991).

3 Evaluation of NLG Models

Several empirical methods have been proposed and applied in the literature for evaluating NLG models. We discuss now why, among the three main evaluation methods (i.e., human judges, corpus-based and task efficacy), task efficacy appears to be the most appropriate for testing the effectiveness of evaluative arguments that are tailored to a complex model of the user's preferences.

The *human judges* evaluation method requires a panel of judges to score outputs of generation models (Chu-Carroll and Carberry 1998; Lester and Porter March 1997). The main limitation of this approach is that the input of the generation process needs to be simple enough to be easily understood by judges[4]. Unfortunately, this is not the case for our argument generator, where the input consists of a possibly complex and novel argument subject (e.g., a new house with a large number of features), and a complex model of the user's preferences.

The *corpus-based* evaluation method (Robin and McKeown 1996) can be applied only when a corpus of input/output pairs is available. A portion of the corpus (the training set) is used to develop a computational model of how the output can be generated from the input. The rest of the corpus (the testing set) is used to evaluate the model. Unfortunately, a corpus for our generator does not exist. Furthermore, it would be difficult and extremely time-consuming to obtain and analyze such a corpus given the complexity of our generator input/output pairs.

[2] This argument was tailored to a default average user, for whom all aspects of a house are equally important. With respect to the first argument, notice the different evaluation for the location and the different order between the two text segments about location and quality.

[3] A threshold controlling verbosity was set to its maximum value.

[4] See (Chu-Carroll and Carberry 1998) 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 dialog system.

When a generator is designed to generate output for users engaged in certain tasks, a natural way to evaluate its effectiveness is by experimenting with users performing those tasks. For instance, in (Young, to appear) different models for generating natural language descriptions of plans are evaluated by measuring how effectively users execute those plans given the descriptions. This evaluation method, called *task efficacy*, allows one to evaluate a generation model without explicitly evaluating its output but by measuring the output's effects on user's behaviors, beliefs and attitudes in the context of the task. The only requirement for this method is the specification of a sensible task.

Task efficacy is the method we have adopted in our evaluation framework.

4 The Evaluation Framework

4.1 The task

Aiming at general results, we chose a rather 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 by considering trade-offs among multiple objectives (e.g., house location, house quality). The selection is performed by evaluating objects with respect to their values for a set of primitive attributes (e.g., house distance form the park, size of the garden). In the evaluation framework we have developed, the user performs this task by using a computer environment (shown in Figure 3) that supports interactive data exploration and analysis (IDEA) (Roth, Chuah et al. 1997). The IDEA environment provides the user with a set of powerful visualization and direct manipulation techniques that facilitate user's autonomous exploration of the set of alternatives and the selection of the preferred alternatives.

Let's examine now how the argument generator, that we described in Section 1, can be evaluated in the context of the selection task, by going through the architecture of the evaluation framework.

4.2 The framework architecture

Figure 2 shows the architecture of the evaluation framework. The framework consists of three main sub-systems: the IDEA system, a User Model Refiner and the Argument Generator. The framework assumes that a model of the user's preferences based on MAUT has been previously acquired using traditional methods from decision theory (Edwards and Barron 1994), 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 the Hot List (see Figure 3 upper right corner) ordered by preference. The IDEA system supports the user in this task (Figure 2 (1)). As the interaction unfolds, all user actions are monitored and collected in the User's Action History (Figure 2 (2a)). Whenever the user feels that she has accomplished the task, the ordered list of preferred alternatives is saved as her Preliminary Decision (Figure 2 (2b)). After that, this list, the User's Action History and the initial Model of User's Preferences are analysed by the User Model Refiner (Figure 2 (3)) to produce a Refined Model of the User's Preferences (Figure 2 (4)).

At this point, the stage is set for argument generation. Given the Refined Model of the User's Preferences for the target selection task, the Argument Generator produces an evaluative argument tailored to the user's model (Figure 2 (5-6)). Finally, the argument is presented to the user by the IDEA system (Figure 2 (7)).

The argument goal is to introduce a new alternative (not included in the dataset initially presented to the user) and to persuade the user that the alternative is worth being considered. The new alternative is designed on the fly to be preferable for the user given her preference model. Once the argument is presented, the user may (a) decide to introduce the new alternative in her Hot List, or (b) decide to further explore the dataset, possibly making changes to the Hot List and introducing the new instance in 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 the new alternative in the first position.

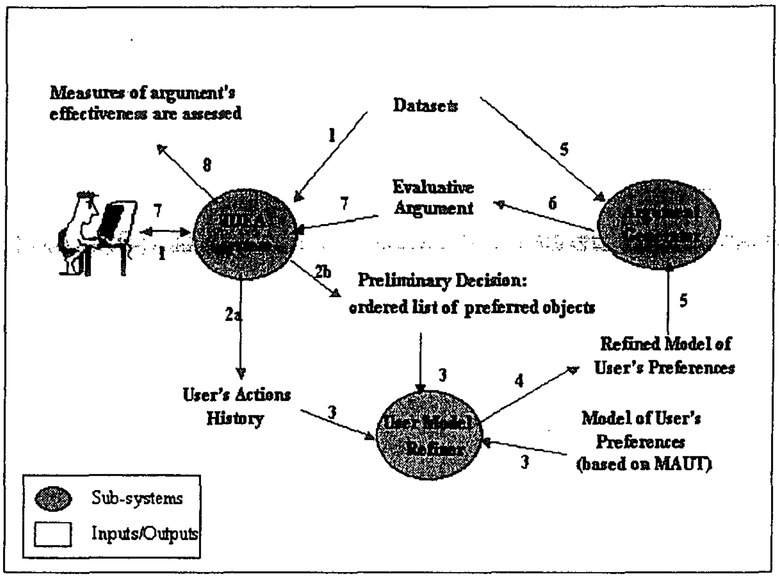

Measures of argument's effectiveness are assessed
Datasets
1
5
8
Evaluative Argument
7
6
7
1
IDEA
2b
Preliminary Decision: ordered list of preferred objects
2a
5
Refined Model of User's Preferences
3
4
User's Actions History
3
User Model Refiner
Model of User's Preferences (based on MAUT)
3
Sub-systems
Inputs/Outputs

Figure 2 The evaluation framework architecture

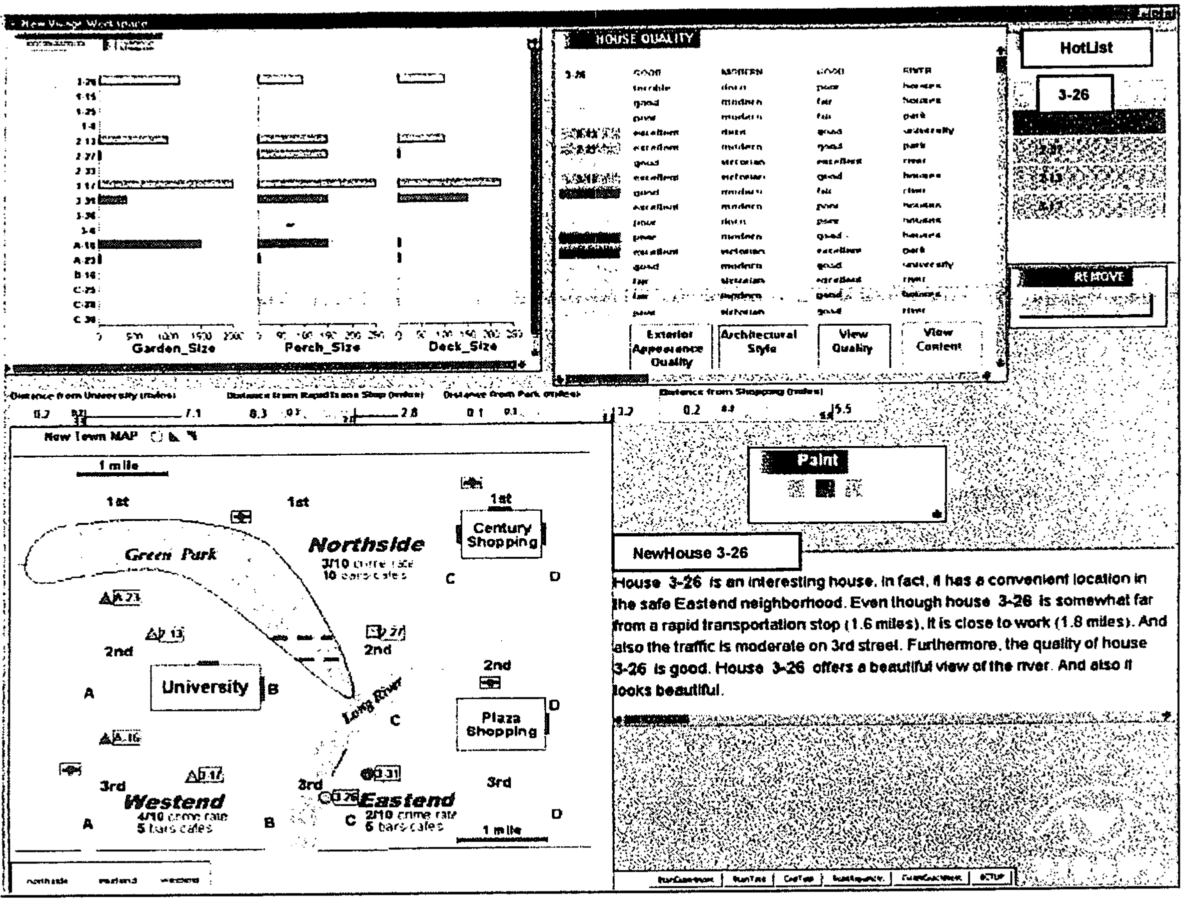
Whenever the user decides to stop exploring and is satisfied and confident with her final selections, measures related to argument's effectiveness can be assessed (Figure 2 (8)). These measures are obtained either from the record of the user interaction with the system or from user self-reports (see Section 2).

First, and most important, are measures of behavioral intentions and attitude change: (a) whether or not the user adopts the new proposed alternative, (b) in which position in the Hot List she places it, (c) how much she likes it, (d) whether or not the user revises the Hot List and (e) how much the user likes the objects in the Hot List. Second, a measure can be obtained of the user's confidence that she has selected the best for her in the set of alternatives. Third, a measure of argument effectiveness can also be derived by explicitly questioning the user at the end of the interaction about the rationale for her decision. This can be done either by asking the user to justify her decision in a written paragraph, or by asking the user to self-report for each attribute of the new house how important the attribute was in her decision (Olso and Zanna 1991). Both methods can provide valuable information on what aspects of the argument were more influential (i.e., better understood and accepted by the user).

A fourth measure of argument effectiveness is to explicitly ask 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. Evaluations based on judgments along these dimensions are clearly weaker than evaluations measuring actual behavioural and attitudinal changes (Olso and Zanna 1991). However, these judgments may provide more information than judgments from independent judges (as in the "human judges" method discussed in Section 3), because they are performed by the addressee of the argument, when the experience of the task is still vivid in her memory.

To summarize, the evaluation framework just described supports users in performing a realistic task at their own pace by interacting with an IDEA system. In the context of this task, an evaluative argument is generated and measurements related to its effectiveness can be performed.

In the next section, we discuss an experiment that we are currently running by using the evaluation framework.



NewHouse 3-26

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.

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

5 The Experiment

As explained in Section 1, the argument generator has been designed to facilitate testing of the effectiveness of different aspects of the generation process. The experimenter can easily control whether the generator tailors the argument to the current user, the degree of conciseness of the argument, and what microplanning tasks are performed. In our initial experiment, because of limited financial and human resources, we focus on the first two aspects for arguments about a single entity. Not because we are not interested in effectiveness of performing microplanning tasks, but because we consider effectiveness of tailoring and conciseness somewhat more difficult, and therefore more interesting to prove.

Thus, we designed a between-subjects experiment with four experimental conditions:

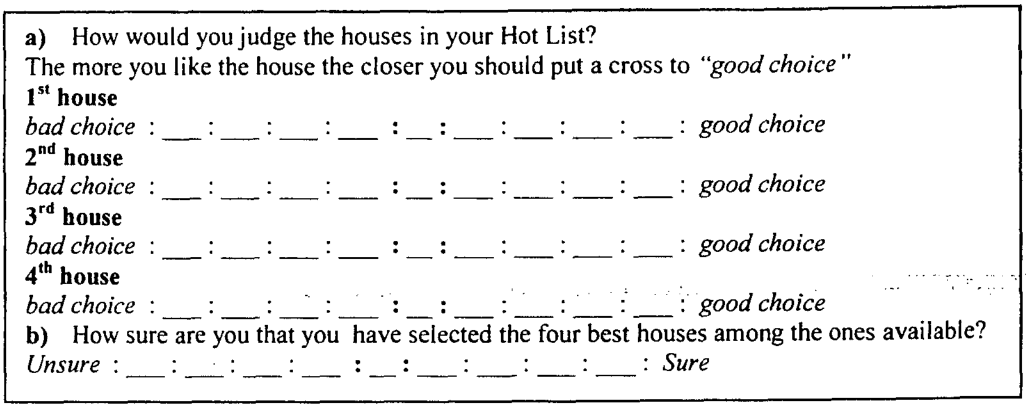
No-Argument - subjects are simply informed that a new house came on the market.

Tailored-Concise - subjects are presented with an evaluation of the new house tailored to their preferences and at a level of conciseness that we hypothesize to be optimal.

Non-Tailored-Concise - subjects are presented with an evaluation of the new house which is not tailored to their preferences[5], but is at a level of conciseness that we hypothesize to be optimal.

Tailored-Verbose - subjects 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.

[5] The evaluative argument is tailored to a default average user, for whom all aspects of a house are equally important.



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 : ___ : ___ : ___ : ___ : __ : ___ : ___ : ___ : ___ : *good choice*

2nd house
bad choice : ___ : ___ : ___ : ___ : __ : ___ : ___ : ___ : ___ : *good choice*

3rd house
bad choice : ___ : ___ : ___ : ___ : __ : ___ : ___ : ___ : ___ : *good choice*

4th house
bad choice : ___ : ___ : ___ : ___ : __ : ___ : ___ : ___ : ___ : *good choice*

b) How sure are you that you have selected the four best houses among the ones available?
Unsure : ___ : ___ : ___ : ___ : __ : ___ : ___ : ___ : ___ : *Sure*

Figure 4 Excerpt from questionnaire that subjects fill out at the end of the interaction

In the four conditions, all the information about the new house is also presented graphically. Our hypotheses on the outcomes of the experiment can be summarized as follows. We expect arguments generated for the Tailored-Concise condition to be more effective than arguments generated for both the Non-Tailored-Concise and Tailored-Verbose conditions. We also expect the Tailored-Concise condition to be somewhat better than the No-Argument condititon, 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 on comparisons of argument effectiveness among the No-Argument, Non-Tailored-Concise and Tailored-Verbose conditions.

The design of our evaluation framework and consequently the design of this experiment take into account that the effectiveness of arguments is determined not only by the argument itself, but also by user's traits such as argumentativeness, need for cognition, self-esteem and intelligence (as described in Section 2). Furthermore, we assume that argument effectiveness can be measured by means of the behavioral intentions and self-reports described in Section 4.2.

The experiment is organized in two phases. In the first phase, the subject fills out three questionnaires on the Web. One questionnaire implements a method from decision theory to acquire a model of the subject's preferences (Edwards and Barron 1994). The second questionnaire assesses the subject's argumentativeness (Infante and Rancer 1982). The last one assesses the subject's need for cognition (Cacioppo, Petty et al. 1984). In the second phase of the experiment, to control for other possible confounding variables (including intelligence and self-esteem), the subject is randomly assigned to one of the four conditions. Then, the subject interacts with the evaluation framework and at the end of the interaction measures of the argument effectiveness are collected. Some details on measures based on subjects' self-reports can be examined in Figure 4, which shows an excerpt from the final questionnaire that subjects are asked to fill out at the end of the interaction.

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

Future Work

In this paper, we propose a task-based framework to evaluate evaluative arguments. We are currently using this framework to run a formal experiment to evaluate arguments about a single entity. However, this is only a first step. The power of the framework is that it enables the design and execution of many different experiments about evaluative arguments. The goal of our current experiment is to verify whether tailoring an evaluative argument to the user and varying the degree of argument conciseness influence argument effectiveness. We envision further experiments along the following lines.

In the short term, we plan to study more complex arguments, including comparisons between two entities, as well as comparisons between mixtures of entities and set of entities. One experiment could assess the influence of tailoring and conciseness on the effectiveness of these more complex arguments. Another possible experiment could compare different argumentative strategies for selecting and organizing the content of these arguments. In the long term, we intend to evaluate techniques to generate evaluative arguments that combine natural language and information graphics (e.g., maps, tables, charts).

Acknowledgements

My thanks go to the members of the Autobrief project: J. Moore, S. Roth, N. Green, S. Kerpedjiev and J. Mattis. I also thank C. Conati for comments on drafts of this paper. This work was supported by grant number DAA-1593K0005 from the Advanced Research Projects Agency (ARPA). Its contents are solely responsibility of the author.

References

Cacioppo, J. T., R. E. Petty, et al. (1984). *The efficient Assessment of need for Cognition.* Journal of Personality Assessment **48**(3): 306-307.

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

Carenini, G. (2000). *Evaluating Multimedia Interactive Arguments in the Context of Data Exploration Tasks.* PhD Thesis, Intelligent System Program, University of Pittsburgh.

Carenini, G. and J. Moore (2000). *A Strategy for Evaluating Evaluative arguments* Int. Conference on NLG, Mitzpe Ramon, Israel.

Chu-Carroll, J. and S. Carberry (1998). *Collaborative Response Generation in Planning Dialogues.* Computational Linguistics **24**(2): 355-400.

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

Dale, R., B. di Eugenio, et al. (1998). *Introduction to the Special Issue on NLG.* Computational Linguistics **24**(3): 345-353.

Edwards, W. and F. H. Barron (1994). *SMARTS and SMARTER: Improved Simple Methods for Multiattribute Utility Measurements.*

Organizational Behavior and Human Decision Processes **60**: 306-325.

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

Eugenio, B. D., J. Moore, et al. (1997). *Learning Features that Predicts Cue Usage.* ACL97, Madrid, Spain.

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

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

Knott, A. (1996). A Data-Driven Methodology for Motivating a Set of Coherence Relations, University of Edinburgh.

Lester, J. C. and B. W. Porter (1997). *Developing and Empirically Evaluating Robust Explanation Generators: The KNIGHT Experiments.* Computational Linguistics **23**(1): 65-101.

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

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

Robin, J. and K. McKeown (1996). *Empirically Designing and Evaluating a New Revision-Based Model for Summary Generation.* Artificial Intelligence Journal, **85**, 135-179.

Roth, S. F., M. C. Chuah, et al. (1997). *Towards an Information Visualization Workspace: Combining Multiple Means of Expression.* Human-Computer Interaction Journal.Vol. 12, No. 1 & 2, pp. 131-185

Shaw, J. (1998). *Clause Aggregation Using Linguistic Knowledge.* 9th Int. Workshop on NLG, Niagara-on-the-Lake, Canada.

Young, M. R. *Using Grice's Maxim of Quantity to Select the Content of Plan Descriptions.* Artificial Intelligence Journal, to appear.

Young, M. R. and J. D. Moore (1994). *Does Discourse Planning Require a Special-Purpose Planner?* Proceedings of the AAAI-94 Workshop on planning for Interagent Communication. Seattle, WA.