Interactive Multimedia Summaries of Evaluative Text

Giuseppe Carenini, Raymond T. Ng and Adam Pauls
Computer Science Department
University of British Columbia
2366 Main Mall, Vancouver, B.C. Canada V6T 1Z4
{carenini,rng,adpauls}@cs.ubc.ca

ABSTRACT

We present an interactive multimedia interface for automatically summarizing large corpora of evaluative text (e.g. online product reviews). We rely on existing techniques for extracting knowledge from the corpora but present a novel approach for conveying that knowledge to the user. Our system presents the extracted knowledge in a hierarchical visualization mode as well as in a natural language summary. We propose a method for reasoning about the extracted knowledge so that the natural language summary can include only the most important information from the corpus. Our approach is interactive in that it allows the user to explore in the original dataset through intuitive visual and textual methods. Results of a formative evaluation of our interface show general satisfaction among users with our approach.

1. INTRODUCTION

Many corporations and organizations are faced with the challenge of managing large corpora of text data. One important application is evaluative text, i.e. any document expressing an evaluation of an entity as either positive or negative. For example, many websites collect large quantities of online customer reviews of consumer electronics. While this literature can be of great strategic value to product designers, planners and manufacturers, the effective processing of this information remains a complex and expensive problem.

An automated solution has the potential of greatly reducing both the cost and time necessary to keep up with the growing amount of review literature. Beyond customer reviews, there are other equally important commercial applications, such as the summarization of travel logs, and non-commercial applications, such as the summarization of candidate reviews. For all these applications, automatic summarization is valuable for managing large amounts of evaluative text. However, in many situations, effectively conveying information about the evaluative text to the user remains problematic. In this paper, we focus on the communication of extracted summaries of evaluative text to the user.

In general, it is widely known that graphics and text are very complementary media with which to effectively convey complex information. While graphics can present large amounts of data compactly and support discovery of trends and relationships, text is much more effective in pointing out and explaining key points about the data, in particular by focusing on specific temporal, causal and evaluative aspects [25]. It is also well known that an effective presentation should always support interactive exploration of the original information source [6, 21]. Thus, to present summaries of evaluative text, our approach is an interactive multimedia one. We aim to convey all the knowledge extracted with graphics, to point out the most important findings in natural language, and also to provide support for exploration of the corpus from which the knowledge was extracted.

To carry out this approach effectively, we need to address three issues. The first is to devise information visualization techniques that can convey all the extracted knowledge to the user. The second is to enable the system to reason about the extracted information so that the most important findings can be intelligently selected for presentation in natural language. Finally, the third issue is to develop a collection of interactive techniques to allow the user to explore the source text corpus in the context of the chosen multimedia presentations. Addressing these three issues and performing a formative evaluation of our approach comprise the key contributions of this paper.

This paper is organized as follows. First, we summarize our approach to knowledge extraction from evaluative text in Section 2. We present the reasoning component of our system that selects the most relevant information for presentation in natural language in Section 3. In Section 4, we describe our visualization technique and discuss how it can effectively convey the extracted information to the user. The interactive multimedia interface which combines visual and textual summaries of the data is described in Section 5. We discuss the results of a formative evaluation of our interface in Section 7. We discuss related work in Section 6, and conclude in Section 8.

2. EXTRACTING KNOWLEDGE FROM EVALUATIVE TEXT

Knowledge extraction from evaluative text about a single entity is typically decomposed in three distinct phases: the determination of features of the entity evaluated in the text, the strength of each evaluation, and the polarity of each

IUI'06, Januaray 29–February 1, 2006, Sydney, Australia.

evaluation. For instance, the information extracted from the sentence *"The menus are very easy to navigate but the user preference dialog is somewhat difficult to locate."* should be that the "menus" and the "user preference dialog" features are evaluated, and that the "menus" receive a very positive evaluation while the "user preference dialog" is evaluated rather negatively.

Our approach to these tasks is described in detail in [5]. While we rely on existing techniques for the second and third tasks of strength and polarity determination [9], we propose for the first task of feature extraction a novel approach that addresses the limitations of earlier work [10]. Specifically, we address the problems of the list of extracted features being too long, scattered, and full of redundancy. Our solution is to automatically map the features extracted by the approach in [10] (called CF for 'crude features') into a user-defined hierarchy of features (called the UDF for 'user-defined features') which describes the entity of interest. See Figure 1 for a sample UDF. Our mapping technique relies on word similarity metrics. Specifically, a crude feature is mapped to a user-defined feature if their mutual similarity exceeds a certain threshold.

Figure 1: Partial view of UDF taxonomies for a digital camera.

Camera	Image
Lens	Image Type
Digital Zoom	TIFF
Optical Zoom	JPEG
...	...
Editing/Viewing	Resolution
Viewfinder	Effective Pixels
...	Aspect Ratio
Flash	...
...	

In [5], we show that the resulting mapping reduces redundancy and provides conceptual organization of the CF. Useful knowledge can be generated through aggregation of the information by UDF, polarity and strength. In this paper, we show that this hierarchical organization can also support the creation of an effective interactive multimedia summary of the extracted knowledge.

3. NATURAL LANGUAGE SUMMARIZATION OF EXTRACTED KNOWLEDGE

As stated in the introduction, our approach to summarizing evaluative text is to convey all the information extracted from the corpus visually while highlighting the most important findings in natural language. In this section we focus on the natural language component of our system whose purpose is to intelligently select the most relevant information for presentation to the user. We approach this reasoning task by defining a 'measure of importance' for each node in the hierarchical set of features and a corresponding procedure for selecting nodes for inclusion into the final summary.

3.1 Selection of Relevant Content

We introduce here the formal definition of our measure of importance. This requires some definitions.

For a corpus of reviews, there is a set of extracted crude features

$$CF = \{cf_j\} \; j = 1...n$$

For example, crude features for a digital camera might include "picture quality", "viewfinder", and "lens". There is also a hierarchical set of user-defined features

$$UDF = \{udf_i\} \; i = 1...m \text{ (cf. Figure 1)}$$

The process of hierarchically organizing the extracted produces a mapping from CF to UDF features. We call the set of crude features mapped to the user-defined feature udf_i $map(udf_i)$. For example, the crude features "unresponsiveness", "delay", and "lag time" would all be mapped to the user-defined feature "delay between shots".

For each cf_j, there is a set of polarity and strength evaluations $ps(cf_j)$ corresponding to each evaluation of cf_j in the corpus. Each polarity and strength evaluation is an integer in the range $[-3, -2, -1, +1, +2, +3]$ where $+3$ is the most positive possible evaluation and -3 is the most negative possible evaluation. We call the set of polarity/strength evaluations directly associated with udf_i

$$PS_i = \bigcup_{cf_j \epsilon map(udf_i)} ps(cf_j)$$

We define the direct measure of importance for a node to be

$$dir\_moi(udf_i) = \sum_{ps_k \epsilon PS_i} |ps_k|^2$$

where by 'direct' we mean the importance derived only from that node and not from its children. The basic premise of this metric is that a feature's importance is proportional to the number of evaluations of that feature in the corpus. However, it seems reasonable that stronger evaluations should be given more weight in the measure of importance than weaker ones. That is, a single evaluation of a feature with a polarity/strength of ± 3 should contribute more to the importance of a feature than an evaluation of ± 1 or ± 2. The sum of squares used for $dir\_moi(udf_i)$ accomplishes both of these goals because it is increased by the number of evaluations, but weighted heavily towards stronger evaluations.

This 'direct' measure of importance, however, is incomplete, as each non-leaf node in the feature hierarchy effectively serves a dual purpose. It is both a feature upon which a user might comment and a category for grouping its subfeatures. Thus, a non-leaf node should be important if either its children are important or the node itself is important. To this end, we have defined the total measure of important $moi(udf_i)$ as

$$moi(udf_i) = \begin{cases} dir\_moi(udf_i) & \text{if } ch(udf_i) = \emptyset \\ [\alpha \; dir\_moi(udf_i) + & \\ (1-\alpha) \sum_{udf_k \epsilon ch(udf_i)} moi(udf_k)] & \text{otherwise} \end{cases}$$

where $ch(udf_i)$ refers to the children of udf_i in the hierarchy and α is some real parameter in the range $[0.5, 1]$. In this measure, the importance of a node is a combination of its direct importance and of the importance of its children. The parameter α may be adjusted to vary the relative

weight of the parent and children. Setting $\alpha = 0.5$ would weight a node equally with its children, while $\alpha = 1.0$ would ignore the importance of the children. We used $\alpha = 0.9$ for our experiments. This setting resulted in more informative summaries during system development.

Finally, we must also define a selection procedure based on this metric. The most obvious is a simple greedy selection – sort the nodes in the UDF by the measure of importance and select the most important node until a desired number of features is included. However, because a node derives part of its 'importance' from its children, it is possible for a node's importance to be dominated by one or more of its children. Including both the child and parent node would be redundant because most of the information is contained in the child. We thus choose a dynamic greedy selection algorithm in which we recalculate the importance of each node after each round of selection, with all previously selected nodes removed from the tree. In this way, if a node that dominates its parent's importance is selected, its parent's importance will be reduced during later rounds of selection. This approach mimics the behaviour of several sentence extraction-based summarizers (e.g. [23, 22]) which define a metric for sentence importance and then greedily select the sentence which minimizes similarity with already selected sentences and maximizes informativeness.

3.2 Generating a Natural Language Summary

Once the most relevant content has been selected, the automatic generation of a natural language summary involves the following additional tasks [20]: (i) structuring the content by ordering and grouping the selected content elements as well as by specifying discourse relations (e.g., supporting vs. opposing evidence) between the resulting groups; (ii) microplanning, which involves lexical selection and sentence planning; and (iii) sentence realization, which produces English text from the output of the microplanner. For all these tasks, we have adapted the Generator of Evaluative Arguments (GEA) [4], a framework for generating user tailored evaluative arguments. Our adaptation of GEA is not the focus of this paper, so we will only provide a brief description here. GEA tailors evaluative arguments about a given entity to a quantitative model of the user preferences that is very similar to the UDF, as it is also describes the entity as a hierarchy of features. Our adaptation relies on this key similarity. In essence, GEA organizes and realizes the selected content as text by applying a strategy based on argumentation theory [3] that considers the strength and polarity of the user evaluation of each feature represented in the user model. Our summarizer applies the same strategy to organize and realize the selected content. However, instead of using the strength and polarity of a user evaluation of each feature, it uses the number of evaluations and an aggregate of the customers opinions about each feature respectively. This aggregate is a function similar in form to the measure of importance used for content selection. For an illustration of the kind of summaries we generate, see the left panel of Figure 4.

4. TREEMAPS FOR PRESENTING THE EXTRACTED KNOWLEDGE

Natural language is very effective in conveying the selected information to the user. However, if we want to communicate all of knowledge extracted by our methodology, graphics will be much more effective at communicating such large amounts of information. We describe in this section the visualization component of our interactive multimedia summary.

For our purposes, an effective visualization technique should

1. Convey the user-defined hierarchical organization of the extracted knowledge
2. Communicate both the importance of and the customer opinions about the extracted knowledge to the user
3. Allow the user to explore the original dataset

We found that Treemaps [24] could be adapted to fulfill all three criteria. A Treemap is a two-dimensional space-filling technique for visualizing hierarchies. A Treemap represents an individual node in a tree as a rectangle with nested rectangles representing the descendants of the node. Because Treemaps use rectangles to represent trees, they can simultaneously visualize the hierarchy (our first criterion) and rapidly communicate other domain-specific information about each node by varying the size and fill color of the rectangles. These two dimensions can be naturally mapped into our domain: size can be used to represent the importance of a feature in the UDF while color can be used to represent customer opinions about a feature. This successfully fulfills the second criterion listed above.

More specifically, to represent the customer opinions for a feature, we used the average of polarity/strength evaluations to set the color a node. Formally, this quantity is

$$av(udf_i) \;=\; \frac{1}{|PS_i|} \sum_{ps_k \epsilon PS_i} ps_k$$

This average was mapped onto a spectrum from bright red for very negative opinions to bright green for very positive opinions. The spectrum grew darker towards the middle such that neutral opinions were entirely black.

Note, however, we do *not* use the measure of importance defined in Section 3 to set the size of a node in the Treemap. Rather, the area of a node in the Treemap is proportional to the number of evaluations of the feature represented by the node

$$count(udf_i) \;=\; |PS_i|$$

We did this for two reasons: firstly, it is critical for the visualization to depend on quantities immediately obvious to the user. Secondly, in some sense, the measure of importance defined earlier for the natural language summary is meant to capture two dimensions (number and strength of evaluations) in a scalar value. Because we are free to represent these two dimensions separately in the Treemap, there is no need to combine them into a less intuitive quantity. Thus, the size of a feature node represents only the frequency of evaluations, while the colour of a feature node represents only strength and polarity of evaluations. See Figure 2 for an example.

It was necessary to make two modifications to a basic Treemap to fully visualize our data. Firstly, Treemaps are generally used for hierarchies in which the descendants of a node form a partition of the node. This is not the case in

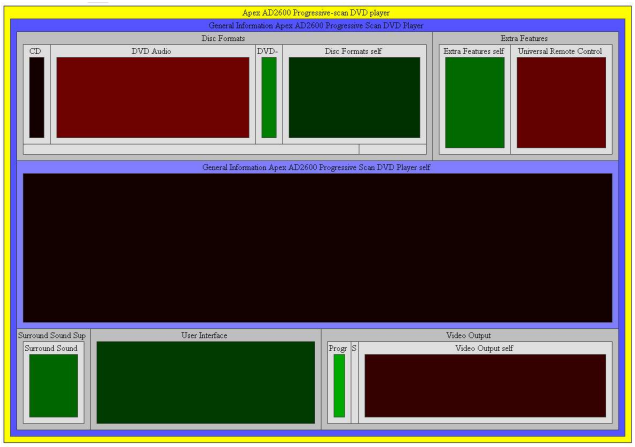

Figure 2: A screen shot of the a Treemap representing the knowledge extracted from a corpus of DVD player reviews.

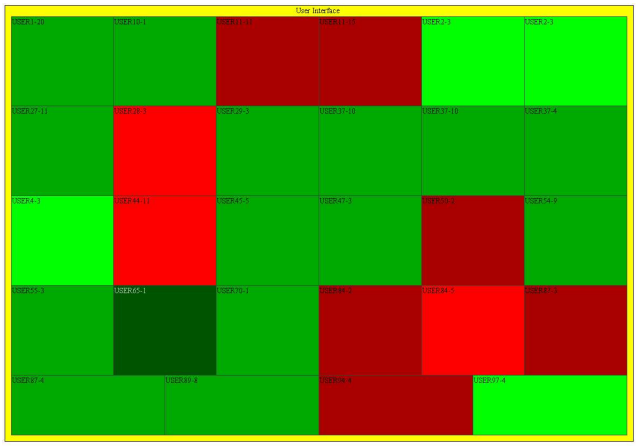

Figure 3: A feature node decomposed into its evaluations in the Treemap interface.

our domain: as mentioned in Section 3, each node in the UDF serves not only as a unifying category for its children, but also as a feature to be evaluated. We thus needed a way to represent the evaluations of a non-leaf node alongside its children. To that end, we created a 'self' node as an additional child for non-leaf nodes which represented the frequency and opinions for the feature represented by the node.

Secondly, we needed to adapt the Treemap to allow exploration of the dataset in order to fulfill the third of our criteria for visualization. We modified the Treemap so that a single node could be decomposed into all of the evaluations of the node. Each evaluation was represented as a rectangle of equal size, and coloured according to its polarity/strength measure. The decomposition of a node can be seen in Figure 3. This decomposition has the dual advantage of allowing the user to see the composition of evaluations which was otherwise averaged into a single color, while also serving as a map from an evaluation back to the original sentence from which the evaluation was extracted.

In terms of implementation, we used the University of Maryland's Treemap 4.1.1 [19] as the basis for our modifications. The source was kindly provided by the authors.

5. INTERACTIVE MULTIMEDIA TREEMAPS FOR SUMMARIZING EVALUATIVE TEXT

We have thus far described the static components of our multimedia summary. As mentioned in the introduction, a good summary of evaluative text should not only present extracted information to the user, but also allow her to explore the original corpus. We present here the interactive methods we have devised to allow the user to do so.

We designed our interactive multimedia interface with the following goals:

1. The user should see the textual summary first. It provides a careful selection of relevant information and also serves to orient the user to the task of examining extracted knowledge.

2. Most of the screen real estate should be devoted to the visualization. Not only does the visualization technique require it, but in all likelihood users will spend most of their time looking at the visualization in order to explore the data.

3. The user should be able to explore the original corpus while still viewing both the summary and the visualization. This should enable the user to quickly verify what s/he is seeing in the multimedia summary.

Figure 4 shows a screenshot of the interface we created with these goals in mind. In the upper left part of the screen, the user sees the textual summary. A Treemap visualization occupies the majority of the upper part of the screen. The bottom of the screen provides space for the user to interactively access the text of the original reviews.

The original set of reviews can be accessed in two ways. Firstly, for each feature evaluated in the textual summary, we provide a small number (usually one or two) of footnotes which point back to reviews which contain sentences which contributed to the evaluation in the summary. When one of these footnotes is clicked, the entire review appears in the bottom section of the screen with the sentence of interest highlighted.

The second way is via the Treemap. The user originally sees the whole Treemap, with the details about individual evaluations hidden. The user can then 'zoom in' to any node to view it in more detail using mouse clicks. 'Zooming in' to a node makes it the root of the tree on the screen. It also decomposes the node into all of its evaluations as described in the previous section (see Figure 3). The user can then click on an evaluation and see the original review from which the evaluation was extracted. As with the footnotes, the sentence from which the evaluation was extracted is highlighted.

6. RELATED WORK

Since text and information graphics are such complementary media it is not surprising that several projects have investigated how they can be effectively integrated in intelligent interfaces (e.g., [6, 13, 26]). However, all these projects differ form our approach for three key reasons. First, they

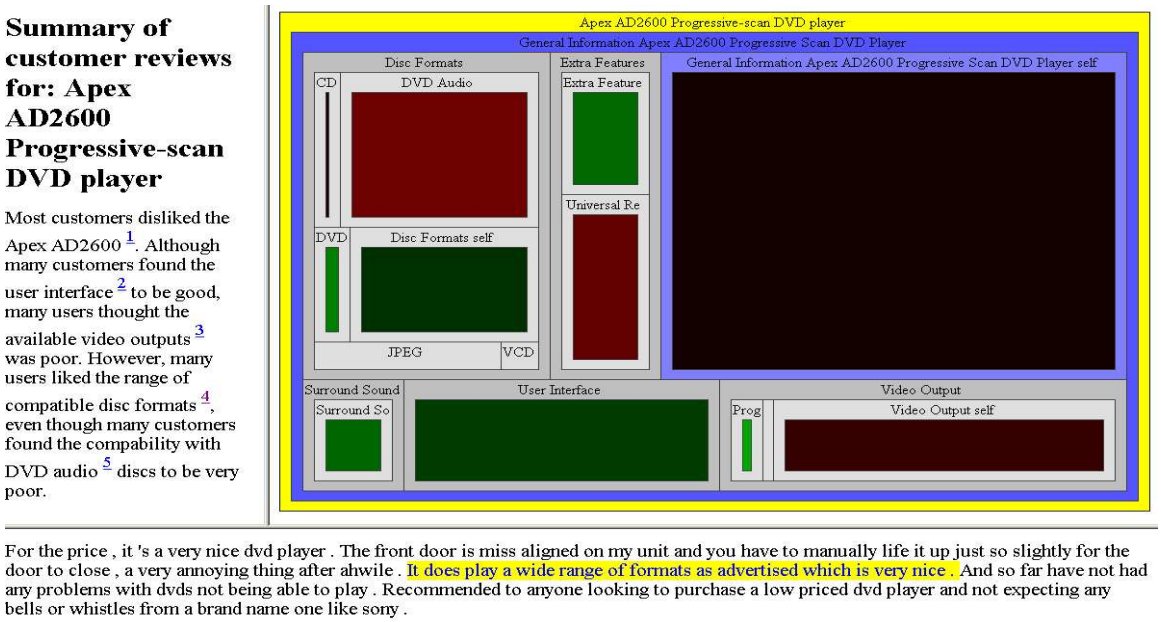

Figure 4: A screen shot of the interface to our interactive summarizer. Each evaluation in the summary corresponds to a node in the Treemap, for example, "available video outputs" refers to the (non-leaf) node in the lower right corner. In this image, the user has clicked on footnote 4, pointing her to a review in which the range of compatible discs is positively evaluated. The text of the review is shown in the bottom of the screen and the relevant sentence is highlighted. This interface can be accessed at `http://www.cs.ubc.ca/~carenini/storage/SEA/demo.html`.

present factual information, while we deal with evaluative information. Secondly, they were limited in using standard/common info graphics or diagrams (e.g., charts, maps) while we have explored integrating text with a quite novel visualization (i.e., Treemaps). The third difference is in the input to the generation process. While previous work generates multimedia given structured data as input, our input, when the whole approach is considered, is a set of documents.

The line of research presented in [1] however may appear to be an exception to this last statement, as it presents a system that generates multi-document summaries integrating text and graphics. Yet, the nature of the generated text and graphics is very different from ours. As is common for research on multidocument summarization [14], text is generated by selecting informative sentences from the source, not by generating language from extracted knowledge. The graphics used are also rather different as the graphical elements displayed do not correspond to extracted information but to whole documents. Specifically, the graphics shows each document as a dot in several semantic spaces and the dot position in those spaces is intended to convey the meaning of the document.

The interest of large organizations in mining online customer reviews and other critical corpora of evaluative text has stimulated considerable research on extracting people's opinions from evaluative text and visualizing the extracted information. As discussed in Section 2, our approach to knowledge extraction relies on the output of Hu and Liu's system [9, 10], which identifies the set of crude product features and the polarity of the corresponding evaluations. As for effectively conveying the knowledge extracted from customer reviews, research has focused on using information visualization exclusively. [12] presents a system that generates a graphical summary of the temporal evolution of customer reviews for a product. For each review, the system simply establishes whether the review is in general positive or negative. Then, since each review has an associated timestamp, a simple stacked barchart can show the proportion of positive and negative reviews for each time-point, therefore supporting the temporal analysis of how these proportions evolve over time (e.g., proportion of negative opinion about a product can drastically decrease once the product is released). With respect to our approach, their knowledge extraction phase and the visualization technique are much simpler than ours, however the temporal aspect of summarizing customer reviews and its visualization is an interesting one that we may consider investigating in future work. Another approach to visualizing information extracted from customer reviews is described in [15]. This work is, however, quite different from ours because their goal is to visualize a comparison among several products of the same type based on an analysis of a corpus of reviews about those products.

7. EVALUATION

We have performed a formative evaluation of our approach to multimedia summarization. The goal was to assess the user's perceived effectiveness of our proposed combination of text and graphics and associated interactive techniques for the task of summarizing large amount of evaluative text.

The focus of this experiment was on the aspects of our approach presented in this paper including (i) the information content of text and graphics in terms of accuracy, precision, and recall; (ii) the integration of text and graphics in terms of redundancy and mutual support; and (iii) the interactive techniques. We did not focus on testing specific aspects of Treemaps, nor of the natural language summarizer.

7.1 The Experiment

Eighteen undergraduate students recruited via an online user experiment system participated in our experiment. A participant was given a brief scripted introduction to Treemaps and allowed to familiarize herself with a sample Treemap. The participant was then given a set of 20 customer reviews randomly selected from a corpus of reviews. Half of the participants received reviews from a corpus of 46 reviews of the Canon G3 digital camera and half received them from a corpus of 101 reviews of the Apex 2600 Progressive Scan DVD player, both obtained from Hu and Liu [8]. The reviews from these corpora which serve as input to our system have been manually annotated with crude features, strength, and polarity. We used a 'gold standard' for crude feature, strength, and polarity extraction because we wanted our experiments to focus on our interface and not be confounded by errors in the knowledge extraction phase.

The participant was told to pretend that they work for the manufacturer of the product (either Canon or Apex). They were told that they would have to provide a 100 word summary of the reviews to the marketing department. The purpose of these instructions was to prime the user to the task of looking for information worthy of summarization. They were then given 20 minutes to explore the set of reviews. The participant could access the reviews through a hypertext interface in which any review can be accessed by clicking on its title. During this time, the participant was allowed to take notes on paper or in a text editor on the computer.

After 20 minutes, the participant was asked to stop. The participant was then given a set of instructions which explained that the company was testing a computer-based system for automatically generating a summary of the reviews s/he has been reading. S/he was then shown the interactive summary generated by our system and given a written explanation of the information displayed by the Treemap and of the associated interactive techniques. The participant was then asked to examine and explore the interactive multimedia summary. Once finished, the participant was asked to fill out a questionnaire assessing the summary along several dimensions related to its effectiveness. The participant could still access the summary while she works on the questionnaire.

Since we could not find in the literature a standard questionnaire specifically designed to assess the effectiveness of multimedia summaries integrating text and graphics, we based our questionnaire on questionnaires developed in previous work for similar assessments including: (i) a questionnaire to acquire feedback from human judges on the effectiveness of multimedia presentations [7]; (ii) questionnaires developed by the NLP community to assess the effectiveness of natural language summaries [16] as well as the effectiveness of natural language advice generated by a multimodal dialog system [11]; and(iii) generic questionnaires for HCI usability testing (especially for the interactive techniques).

Question	Average	Standard Deviation
Structure	4.35	0.61
Attractiveness	3.65	0.86
Recall	4.41	0.94
Precision	4.14	1.03
Accuracy	3.82	0.95
Redundancy	4.40	0.63
Text summary	3.71	1.16
Text support	4.24	1.09

Table 1: Quantative results of user responses to our questionnaire on a scale from 1 (Strongly Disagree) to 5 (Strongly Agree). See Appendiex 8 for the exact wording of the questions.

The questionnaire provided to the participants can be found in Appendix A.

7.2 Results

In this section, we briefly discuss the quantitative results of our experiment. We then move on to discuss the most important qualitative comments received from users and how we plan to redesign our system to address these comments.

7.2.1 Quantitative Results

The quantitative results are summarized in Table 1. One of the 18 subjects misread the instructions so his results were not used. There were also three responses to Question 6 which indicated a misunderstanding of the question (participants commented positively about the summarizer but circled either "Disagree" or "Strongly Disagree"). These individual responses were not used.

Overall, the participants seemed to be happy with the interactive summary. The only categories under which responses averaged under an "Agree" level were Attractiveness, Accuracy, and Text Summary. We discuss participant complaints with the physical appearance of the interface along with the qualitative results. With regard to Accuracy, we are not surprised that the summarizer scored low. The process of mapping crude features into the UDF frequently creates spurious mappings (e.g. "DVD Audio" to "Audio Quality"). However, only 2 users noticed such incorrect mappings. Comments indicated that most users who gave a low score for Accuracy were concerned that the summarizer missed more subtle details (e.g. "*Some comments are mixed reviews which is hard to reflect in the summary.*").

With regard to the textual summary, we are also not surprised that it scored (relatively) low. Automated summarization is a hard task, and most mature systems are not human-competitive yet [17]. We note, however, that the average responses for Text Support were quite high, indicating that even though the textual summary was not very good on its own, participants felt that it served as an effective overview of the information in the corpus when integrated with the graphical information. One user stated this explicitly: "*I can get a general idea from the text and detailed information from the graphics.*"

7.2.2 Qualitative Results

We take the generally positive quantitative response to indicate that our interface is a promising approach that may require mostly minor tweaking to satisfy most users. Addi-

tionally, users were in general quite pleased with the interactive elements of our interface (e.g. "*The links to the actual passages are useful for referencing and going into more depth about the comment where needed.*", ' '*I like the highlight part! [sic]*"). In this section, we focus on the most common user criticisms and how we could redesign our interactive multimedia summary to meet these needs.

Criticisms of the multimedia summary fell largely into the following categories:

1. Complaints about the physical appearance of the Treemap and the insufficient size of many of the rectangles.
2. Difficulty understanding the Treemap representation or manipulating it.
3. A strong preference for the textual summary over the graphical representation or vice versa.
4. Confusion about the content and ordering of evaluations in the textual summary or how the textual summary related to the visualization.

· With respect to (1), most problems (such as the font size and color selection) are easily fixable. Complaints about the insufficient size of some rectangles requires a little more thought to address. This problem occurs when the size of a rectangle is large enough to be shown on the Treemap, but so small that the border around it and the associated text is bigger than the rectangle. The simplest solution is probably to have a larger minimum size for rectangles, such that smaller rectangles are not shown even if there is room on the screen.

With respect to (2), there is evidence that Treemaps can be difficult for novice users to understand [18]. However, [18] also identifies several recent real-world applications of Treemaps which have been successful even for untrained users. In each case, success was contigent on (i) the users being familiar with the visualized data and (ii) extensive user interface refinement. In a real-world application of our interface, (i) would be less of a problem because users would generally be more familiar with the summarized entity because they would have chosen to view it (rather than having the entity thrust upon them in a laboratory setting). To further refine the user interface, more user studies would be required. We note, however, that the majority of participants had little trouble grasping the visualization. As such, we believe that Treemaps provide a viable method for the visualization task at hand.

Comments relating to (3) were actually very encouraging. Only two users explicitly stated that they preferred the visual summary over the textual summary, while a total of five participants expressed a strong preference for the textual summary over the visualization. This appears to support our assumption that an effective summarization technique requires both natural language and visualization to be effective for all users.

Most users seemed to spend most of their time exploring the Treemap, and tended not to look at the textual summary for very long. As such, comments relating to (4) were relatively few. However, one somewhat frequent complaint was the perception that the evaluations in the summary were poorly organized. Participants suggested that positive and negative comments be grouped together rather than listed in seemingly random order as they were. The ordering of features in the textual summary is entirely based on a depth-first traversal of the hierarchy, which is intended to provide coherence in the summary. It appears, then, to have the opposite effect in some cases. Notice however that the structuring of the summary content is not the focus of this paper so we will not discuss this any further.

Another complaint about the textual summary was that some users did not immediately see how evaluations in the text corresponded to nodes in the Treemap. One wanted the Treemap reorganized such that the order of evaluations in the text matched the order in the Treemap, while others wanted explicit use of the Treemap labels in the text. The first suggestion could be easily implemented, while the second would be best addressed by better coordination of text and graphics. We plan to address this problem in future work discussed in the next section.

8. CONCLUSIONS AND FUTURE WORK

We have presented a multimedia interactive technique for summarizing a set of evaluative arguments about an entity. This technique utilizes Treemaps to visualize all the knowledge extracted from the corpus and natural language to create an overview of the most relevant information. Both the text and the Treemap permit exploration of the original dataset. We have evaluated our approach in the domain of online reviews of consumer electronics. A formative user study of our summarization interface indicates that users found it both intuitive and informative. Negative feedback about the interface focussed largely on the superficial aspects of the interface, and so we regard the core functionality of the summarizer as successful.

In the future, we plan to continue our investigation along the following lines. In our multimedia summaries, text and graphics are integrated in the sense that text supports the graphics by highlighting the most important findings. However, text and graphics are not coordinated in any explicit way, which confused some users as we explained in the previous section. One potential solution is this: when the user is exploring information in the Treemap, if the same information has also been selected in the textual summary, the system could highlight the corresponding sentence in the summary. This coordination might also address another problem we noticed in the formative evaluation, namely that some users tended to pay less attention to the textual summary.

From our experiments it appears that Treemaps are quite effective in conveying the hierarchically structured evaluations extracted from a corpus of customer reviews. However, alternative visualization techniques have been recently proposed for a similar task [2]. So, we intend to verify whether this novel techniques may be more effective than treemaps in supporting the integration of text and graphics in our multimedia summaries.

9. REFERENCES

[1] K. R. Ando, B. K. Boguraev, R. J. Byrd, and M. Neff. Multi-document summarization by visualizing topical content. In *Proc. of ANLP-NAACL Workshop on Automatic Summarization*, 2000.

[2] D. Brodbeck and L. Girardin. Visualization of large-scale customer satisfacyion surveys using a parallel coordinate tree. In *Proc. of the IEEE Symposium on Information Visualization*, 2003.

[3] G. Carenini and J. D. Moore. A strategy for generating evaluative arguments. In *First International Conference on Natural Language Generation*, pages 47–54, Mitzpe Ramon, Israel, 2000.

[4] G. Carenini and J. D. Moore. An empirical study of the influence of user tailoring on evaluative argument effectiveness. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Seattle, USA,, 2001.

[5] G. Carenini, R. Ng, and E. Zwart. Extracting knowlede from evaluative text. In *Proc. Third International Conference on Knowledge Capture*, 2005.

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

[7] H. Guo and A. J. Stent. Trainable adaptable multimedia presentation generation. In *Proceedings of the International conferences on Multimodal Interfaces*, Trento, Italy, 2005.

[8] M. Hu and B. Liu. Feature based summary of customer reviews dataset. *http://www.cs.uic.edu/ liub/FBS/FBS.html*, 2004.

[9] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proc. ACM SIGKDD*, 2004.

[10] M. Hu and B. Liu. Mining opinion features in customer reviews. In *Proc. AAAI*, 2004.

[11] M. Johnston, P. Ehlen, S. Bangalore, M. Walker, A. Stent, P. Maloor, and S. Whittaker. Match: An architecture for multimodal dialogue systems. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 376–383, Philadelphia, USA, 2002.

[12] D. Kusui, K. Tateishi, and T. Fukoshima. Information extraction and visualization from internet documents. *NEC Journal of Advanced technology*, 2(2):157–163, 2005.

[13] M. F. G. Lapalme. Intentions in the coordinated generation of graphics and text from tabular data. *Knowledge and Information Systems*, 2:3, 2000.

[14] I. Mani and M. Maybury, editors. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, Ma., 1999.

[15] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining product reputations on the web. In *Proc. ACM SIGKDD*, 2002.

[16] Guidelines of the 2004 document understanding conference. *http://www-nlpir.nist.gov/projects/duc/guidelines/2004.html*, 2004.

[17] Online proceedings of the 2004 document understanding conference. *http://duc.nist.gov/pubs.html#2004*, 2004.

[18] C. Plaisant. The challenge of information visualization evaluation. In *AVI '04: Proceedings of the working conference on Advanced visual interfaces*, pages 109–116, New York, NY, USA, 2004. ACM Press.

[19] C. Plaisant, B. Shneiderman, G. Chintalapani, and A. Aris. Treemap home page. *http://www.cs.umd.edu/hcil/treemap/*.

[20] E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, 2000.

[21] S. F. Roth, P. Lucas, J. A. Senn, C. C. Gomberg, M. B. Burks, P. J. Stroffolino, J. A. Kolojejchick, and C. Dunmire. Visage: A user interface environment for exploring information. In *Proceedings of Information Visualization*, pages 3–12, San Francisco, Ca., 1996.

[22] H. Saggion and R. Gaizauskas. Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of Document Understanding Conference DUC04*, 2004.

[23] B. Schiffman, A. Nenkova, and K. McKeown. Experiments in multidocument summarization. In *Proceedings of Human Language Technology HLT02*, San Diego, Ca., 2002.

[24] B. Shneiderman. Tree visualization with treemaps: A 2-d space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99, 1992.

[25] E. R. Tufte. *Visual Explanations: Images and Quantities, Evidence and Narrative*. 1997.

[26] J. Yu, E. Reiter, J. Hunter, and C. Mellish. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, to appear, 2005.

Appendix A: Questionnaire used in the the Formative User Study

The questionnaire presented to the participants included the following statements. Participants indicated their degree of agreement on the standard 5 point Likert scale (one of "Strongly Disagree", "Disagree", "Neutral", "Agree", "Strongly Agree" or "No Opinion"). The participant was provided with space for free text comments after each question.

1. *This multimedia summary (i.e. both text and graphics) is clear, well-structured and well organized.* (Structure)
2. *This multimedia summary is attractive.* (Attractiveness)
3. *This multimedia summary contains all of the information you would have included from the source text.* (Recall)
4. *The multimedia summary does not contain information from the source text that you would have left out (i.e. does not contain unneeded/extra information).* (Precision)
5. *All information expressed in the multimedia summary accurately reflects the information contained in the source text.* (Accuracy)
6. *In the multimedia summary, text and graphics present the right degree of redundancy.* (Redundancy)
7. *The textual summary was a good summary of the source text.* (Text summary)
8. *In the multimedia summary the text summary effectively highlight the most important information displayed by the graphics.* (Text support)

In addition, the participant was also asked several open-ended questions about his/her interaction with the multimedia summary.

- *What did you like about the summary interactive techniques?*
- *What did you not like about the summary interactive techniques?*
- *What did you find confusing in the summary interactive techniques?*
- *How would you suggest improving the summary interactive techniques?*