Improving Meeting Summarization by Focusing on User Needs: A Task-Oriented Evaluation

Pei-Yun Hsueh  
School of Informatics  
University of Edinburgh  
pshsueh@us.ibm.com

Johanna D. Moore  
School of Informatics  
University of Edinburgh  
J.Moore@ed.ac.uk

ABSTRACT
Advances in multimedia technologies have enabled the creation of huge archives of audio-video recordings of meetings, and there is burgeoning interest in developing meeting browsers to help users better leverage these archives. A recent study has shown that extractive summaries provide a more efficient way of navigating meeting content than simply reading through the transcript and using the audio-video record, or navigating via keyword search [15]. The extractive summary technique identifies informative dialogue acts to generate general purpose summaries. These summaries can still be lengthy. Recently, we have developed a decision-focused summarization system that presents only 1-2% of the recordings related to decision making. In this paper, we describe a task-based evaluation in which we compare the decision-focused summaries to the general purpose summaries. Our results indicate that the more focused summaries help users perform the decision debriefing task more effectively and improve perceived efficiency. In addition, this study also investigates the effect of automatic summaries and transcription on task effectiveness, report quality, and users’ perceptions of task success.

Author Keywords
Meeting browser, automatic summarization, multimedia information retrieval, task-oriented evaluation, user study.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: Miscellaneous—Optional sub-category

INTRODUCTION
Meetings are a critical aspect of most organizations. In meetings two or more people gather to discuss a topic, hoping to reach conclusions through the communication process. This process involves a shared goal and intensive oral arguments which provide rationales for individuals’ points of view. Repositories of the audio-visual recordings of meetings two or more people gather to discuss a topic, hoping to reach conclusions through the communication process. This process involves a shared goal and intensive oral arguments which provide rationales for individuals’ points of view. Repositories of the audio-visual recordings of meetings constitute a valuable source of information for future training and group decision support [19, 21]. With the recent advances in recording and storage technologies, a rapidly growing number of meetings are being archived for later retrieval, and solutions are needed to help users better leverage the archived meeting recordings.

Standard meeting browsers, which come with typical information retrieval and playback facilities, help answer less than 20% of user queries [18]. This has led researchers to augment meeting browsers with additional plug-ins. For example, plug-ins that display topics, and represent speaker roles and meeting states have been found to be effective for meeting information retrieval, helping users find the information they seek in 25% less time [1].

In addition, organizational and user query studies have been conducted to identify what key aspects are missing from current meeting browsers [4, 11, 18]. Pallotta et al. [18] have highlighted the argumentation process and outcome as the most sought-after information, composing 60% of common user queries. Indeed, it has been suggested that argumentation outcomes, i.e., decisions, are the most essential type of information that users require about meetings [18, 19, 20, 21, 24].

In this paper, we investigate whether augmenting a meeting browser with a query-focused summary will help users find information from the archives more effectively. Specifically, we evaluate the use of two types of summaries for the task of identifying the essential argumentation outcomes from previous meetings, a common practice for meeting preparation [17, 20, 24]. The first type is an extractive summary that provides a general overview of the meeting (hereafter, “general-purpose summary display”), and is the type of summary produced by existing summarization systems. Murray [15] has shown that such summaries are rated highly by users performing a decision audit in which they had to summarize how a group came to a particular decision.

However, because general-purpose summaries are often lengthy—Murray [15] extracted 30-40% of the meeting dialogue acts—they are expected to be less effective for the decision debriefing task than summaries that are focused solely on decisions, which include only 1-2% of the meeting dialogue acts, see [7]. Therefore, in this paper, we also evaluate the use of a display that presents summaries that are tailored to the user task (which we will call the “decision-focused summary display”).
Figure 1. Example AMI browsers. Both are composed of three plug-ins: the playback facility of audio-video recordings (top), the transcription display (lower left), and the extractive summary display (lower right). The two browsers differ only in the summary display plug-in, with the browser on the top showing the general-purpose summary display and the one on the bottom showing the decision-focused summary display.
Figure 1 shows the two types of summaries, along with the transcripts and audio-video recordings in a meeting browser. To the best of our knowledge, no user study has been performed comparing the effectiveness of using general-purpose vs. decision-focused summaries. Moreover, no studies have yet investigated how users can leverage extractive summaries to obtain an overview of all of the decision-relevant information in a meeting in order to produce an abstractive summary of the decisions made in meetings.

RELATED WORK

To save time and human labor in generating meeting minutes, techniques have been developed to produce extractive summaries by distinguishing the informative dialogue units from the uninformative ones in meetings.

Traditionally, the extractive technique works well in text summarization. For example, Mani and Bloedorn [12] found that users absorb information in summaries more quickly than in full text, despite some loss of accuracy. Text summarization commonly uses lexical information, such as counts of cue phrases, word co-occurrences, and tf*idf scores (or its variants), to rank the extract-worthiness of each unit [5, 9, 22]. Some methods rely on orthographic cues (e.g., the position in text, title) and semantic information (e.g., the degree of connectedness in a semantic graph, co-reference) [2, 12]. However, not all of these features are not available for speech. Previous research in speech summarization overcomes this problem by using other types of speech-specific information. For example, lexical and prosodic information have been combined to perform summarization in speech genres such as broadcast news [8] and voice-mail [13].

Extractive techniques have also been applied to meetings in order to identify informative units that are reflective of overall meeting content [14, 16, 25]. For example, [15] used prosodic information to identify the most informative meeting dialogue acts for general purpose extractive summaries. However, although general purpose summaries were shown to be effective for the decision audit task, these typically lengthy summaries may be less effective for users who need a quick overview of all decisions made in a series of meetings. In our own prior work [7], we have adopted a query-driven summarization approach, selecting only those dialogue units that help fill in a decision-related template. The technique developed in this work identified lexical as well as multi-modal features (e.g., gestures, head movements, prosody) that are predictive of decision-related units.

METHODOLOGY

Task overview

Obtaining an overview of the decisions made in previous meetings is critical to the preparation for future meetings [17, 20, 24]. Thus we chose a “decision debriefing” task for this study comparing the two types of extractive summaries. In this task, participants are asked to produce an abstractive summary of all the decisions made in a series of meetings, using hints from the extractive summaries. The abstractive summaries the subjects produce are expected to be easier to interpret. This is because extractive summaries, which are simply a list of dialogue acts extracted from the meeting transcript often seem disjointed or incomplete.

We recruited 35 participants (20 females and 15 males, ages from 18 to 44) during a two-month period in 2008. These subjects were recruited from the undergraduate and graduate program of distinctly diverse fields (e.g., history, medicine, chemistry, geography). Before beginning the task, participants were asked to fill in a questionnaire about their prior experience in computer usage and meeting attendance. An experimenter then guided the subject through the procedure.

Participants were asked to analyze a series of four design meetings (each 30-40 minutes) and summarize the decisions in a report for their upper management. For the reader’s information, the following is a synopsis of the decision summary for the series of meetings used in our experiment. Note that the human authored, gold standard summaries are written in the form of bullet points (as exemplified below in Fig. 2).

- In the kick-off meeting (A), the entire group decided that the prototype design should be simple, keeping the everyday functions on one interface and more complicated functions on another;
- In the conceptual design (B) and detailed design (C) meetings, the group decided on the specific target user group, the essential functions of the interface and the layout;
- In the wrap-up/evaluation meeting (D), the group decided on which prototype to choose and what functions to be eliminated from the prototype.

At the beginning of each session, the experimenter introduced the browser interface to the participant. The participants were then free to browse through one pre-selected meeting recording (which is not used in the real experiment). They could take as much time they needed to familiarize themselves with the interface.

The main task is as follows:

“In 45 minutes or less, write a report to summarize the decisions made in the four meetings for upper management. Our pilot study demonstrated that the decision debriefing task is straightforward enough to be completed in 45 minutes.

Because some participants in a pilot study expressed the need to be reminded of the time remaining in the experiment, the experimenter signalled the participants twice before the end of the experiment, once at 25 minutes and again at 40 minutes into the experiment. The participants could also signal the experimenter to end the session if they finished the task early. During the session, all user behaviors as well as the user-generated decision minutes, were recorded in log files.

At the end of each session, the experimenter asked the participant to explain how they used the browser interface to identify decisions made in the meetings. Participants were also asked to fill in a post-questionnaire about their perceived task success.
**Meeting Corpus and Annotation**

To obtain human-generated decision-focused extracts for use in training our automatic decision extraction model, and decision-focused abstracts to provide a gold standard for evaluating the decision summaries generated by participants in our study, we use the AMI meeting corpus [3], in which meeting participants are required to make decisions as a group in a series of four product design meetings, intended to imitate a typical product design cycle. As part of the collection and annotation of this meeting corpus, a two-phase procedure for generating summaries was performed:

- First, annotators were asked to navigate the recordings of one series of four meetings and to summarize the decisions made in these meetings into a list of “decision points” (as exemplified in Figure 2).
  - The group decided not to define the target user group by a specific age range but simply by interest in fashion and simplicity.
  - The remote will feature a locator function and large buttons.
  - The remote will incorporate both simple and complicated functions, hiding the complicated functions from the main interface.
  - The remote will be made to look fashionable.

**Figure 2. Example decision points of a product design meeting.**

- The annotators were asked to go through the dialogue acts (DAs) in each meeting one by one, and to determine whether the DA supported any of the decision points in the summary. If so, the act was tagged as a decision-related dialogue act (“decision DA”).

**Decision point annotation**

In the first phase of the annotation procedure, the set of decision points that were noted by two or more annotators are used as the gold standard set of decision points. In the meeting series used in this study, the meeting participants reached 6, 10, 8, and 6 decisions respectively.

<table>
<thead>
<tr>
<th>Coder Pair</th>
<th>1,2</th>
<th>1,3</th>
<th>2,3</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa</td>
<td>0.58</td>
<td>0.54</td>
<td>0.70</td>
<td>0.61</td>
</tr>
</tbody>
</table>

**Table 1. Pair-wise agreement of decision-focused extract annotations.**

**Decision-focused extract annotation**

After annotation, the ground truth “decision-focused extract” of each meeting is then generated by collecting the set of decision DAs that were extracted by one or more annotators.

An analysis of the decision-focused extract annotation shows that the annotators found on average four decision points per meeting and specified around two decision-related dialogue acts for each decision sentence in the set of decision points. Overall, 554 out of 37,400 DAs (in a 50 meeting dataset) were annotated as decision DAs, accounting for 1.4% of all DAs in the data set.

To examine the level of inter-coder agreement of the decision DAs, we calculated Kappa statistics $\kappa$ between each pair of annotators’ decision-focused extracts. Table 1 presents the reliability of the annotation procedure. [15] reported the level of agreement in the annotations of the original extractive summaries as 0.48 (average Kappa value).

**Meeting Browser Interface**

The meeting browser (cf. Fig. 1) used in this evaluation is an enhanced version of the J-Ferret meeting browser [23], which is designed to present additional annotations on top of the meeting recordings [3]. The enhanced version consists of three basic components: the audio-visual recording playback facility (top), the transcript display (lower left), and the extractive summary display (lower right).

Each participant is equipped with a headphone so that they can listen to the audio recordings whenever necessary. Users can play the audio-video recording from the beginning. Users who are interested in a particular decision DA can click on that “DA button” in the display and be led to the point where the DA was uttered in the dialogue. Each of the decision DA buttons in the summary display is time synchronized to the location of the decision DA in the audio-visual recording as well as in the transcript.

There are five tabs on the top of the browsing interface: (1) the first four tabs take users to each of the four meetings in the series chosen for display, and (2) the last tab is the “writing tab”, which brings up the page on which users can type in their summaries. Users can switch between these tabs at will. During the experiment, a logging tool in the back-end records all the clicking and typing behaviors. With this log, we can analyze the use of the different components in the browser, such as the summary display and the audio-video playback facility, as well as the report typing behavior, e.g., how many characters were deleted, inserted, and substituted by each participant.

**Automatic decision-focused extracts**

Because our ultimate aim is to generate decision-focused summaries automatically, we also evaluated the effectiveness of the automatic decision-focused extracts. The automatic extracts were generated by a state-of-the-art algorithm of [7], which trained ensemble Maximum Entropy (MaxEnt) models to classify each DA as a decision DA or not. The same number of decision DAs as in the ground truth data were then selected into the set of automatic decision-focused extracts. This is to ensure the length of the automatic extracts are the same as the manual ones so that no effect of summary length would be introduced.

We then checked how many decision points in each meeting would not be captured by carefully navigating through the set of DAs that are displayed with the transcripts and recordings in the browser. For example, in Meeting C which contains eight decisions, this method would miss 3-6 major decisions; A more casual reading of the transcripts may result in the misses of two more decisions and the misinterpretation of one decision. Among the 30 decisions in the recordings, using the automatic DAs would capture 17-23 of them (57-76%), depending on whether the participant reads through the neighboring transcripts or not.
Automatic speech recognized transcription
The ASR transcription used in this experiment was generated with a state-of-the-art program, which on average recognizes words with 30%-40% error rate, that is, a 30%-40% chance for a word to be substituted, deleted, or incorrectly inserted. The ASR system used a vocabulary of 50,000 words, with a trigram language model trained on a combination of in-domain meeting data, related texts found by web search, conversational telephone speech (CTS) transcripts and broadcast news transcripts (about $10^9$ words in total). Additional technical details can be found in [6].

Using ASR transcription rendered the interpretation of some decisions difficult. For example, the DA “the buttons can be fruit shaped” was recognized as “the buttons can be fair share”, and the word “titanium” as “taking”.

Experiment Design
Our hypothesis is that a more succinct, decision-focused summary (a small portion of the general purpose extractive summary) would help users prepare a meeting minute for upper management more effectively and feel more confident about their meeting preparation work. In the context of the decision debriefing task, this study aims to answer the main question:

- Are extractive summaries that focus on decision-related dialogue acts more useful than general-purpose extractive summaries—which extract generically important dialogue acts that reflect overall meeting content—for the decision debriefing task?

In addition, since we would like to know how much automation degrades the usefulness of the summary display, we also address the following two questions:

- Can the automatic decision-focused extracts help users achieve performance comparable to that obtained by navigating the manual extracts?
- Does operating on transcripts produced by automatic speech recognition (ASR), as opposed to manually produced transcripts, significantly affect user performance?

In our study, each participant was randomly assigned to one of four groups. Each group was asked to accomplish the decision debriefing task, using one of the following summary displays embedded in the meeting browser:

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Levels</th>
<th>Dependent Variable</th>
<th>Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extract Type</td>
<td>General-purpose</td>
<td>Task effectiveness (User summary-based)</td>
<td>Ratio of correctly found decisions to all decisions in the model summary</td>
</tr>
<tr>
<td></td>
<td>Decision-focused</td>
<td>Report quality (User summary-based)</td>
<td>Overall quality, completeness, conciseness, trustworthiness, style</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Perceived success (Post-questionnaire)</td>
<td>Ease of use, task completeness, Decision coverage and comprehension</td>
</tr>
</tbody>
</table>

Table 2. Experimental design.

In the Baseline condition, we used what is generated in Murray (2007) [15]. The lengths of the general summaries represent compressions of approximately 40%, 32%, 32% and 30% of the original meeting length. The AD-ASR condition is generated by projecting the position of the automatically identified decision DAs on the ASR transcripts. Both the transcripts and the decision-focused extracts are based on the error-prone ASR outputs, leading to the occasional misinterpretation of what the meeting participants were saying.

We designed our experiment around the hypothesis that a decision-focused summary display benefits users more than a general-purpose display for accomplishing the decision debriefing task. The independent and dependent variables are shown in Table 2. The independent variables are tested between participants.

The dependent variables are classified into three categories:

1. Task effectiveness: First of all, the user-generated decision summary is evaluated against the set of gold standard decision points. Task effectiveness is measured by the percentage of gold standard decision points that are correctly listed in the user’s decision summary ("decision hits"). We followed the tradition of the summarization research community in using recall-based measures, e.g., ROUGE [10]. In addition, because participants assumed different writing styles in the decision debriefing task, the calculation of precision is impractical.

2. Report quality: Several aspects of the user-generated decision summary were rated on 7-point Likert scales by a human judge. These aspects include the overall quality, completeness, conciseness, task comprehension, amount of effort spent in writing, trustworthiness, and writing style. The human judge was given a set of criteria for scoring each of these aspects.
3. User perceived success: Finally, there are the self-reported measures of the level of perceived success and usability, reported on 5-point Likert-scales in the post-questionnaire. We did not use 7-point scales in the questionnaire because our questionnaire is a modified version of the one used in [15] and we wanted to compare the level of perceived success of general and decision-focused extractive summaries. These measures include perceived ease of use, ease of search, efficiency in finding all relevant information, task completeness, general task comprehension, task success, task difficulty, and pressure.

In addition to the three evaluation criteria, we have also developed a number of quantifiable measures to understand user behavior when given the different types of summary displays. These measures are computed from the log files. For example, the usage of summary display is measured by counting the number of decision DAs the user clicks on. The assumption is that the more DAs selected by the user, the more likely it is that the summary presents information that the user wants to know. A more detailed account of the log-based measures of user behaviors is presented in Table 3.

**RESULTS**

Our research questions are concerned with the effect of summary type on users’ task performance and perceived success. In addition, we also test the impact of automation on the performance of the decision-focused summary.

**Effect of Summary Type on Decision Debriefing**

In this section, we report the results obtained from analysis of the log files and the user-generated summaries.

**Task effectiveness analysis**

Our analysis shows that, on average, participants who were given the decision-focused extractive summary display yielded more decision hits than participants given the general-purpose one (Fig. 3). To determine whether the differences are statistically significant, an analysis of variance was performed. The meeting summary display type was found to have a significant main effect on task effectiveness ($F(3, 31) = 13.832; p < 0.001$). The best performing participant was able to use the TOPLINE (manual decision extract, manual transcription) browser to find almost all of the decision points.

In Fig. 3 we also report the participants’ task effectiveness in the first three meetings. By comparing the results obtained from the first three meetings and those from all four meetings, we can examine the effect of the different conditions on task completeness, i.e., how many meetings users can effectively analyze in order to find decisions in the time available. The comparison shows that the above result holds true when the decision hits are measured in the first three meetings only, indicating that there is not a major impact of the conditions on task completeness in this experimental design.

**Report quality analysis**

A condition (4) x report quality (5) analysis of variance on the decision minute ratings (Table 4) finds the meeting summary display type to also have a significant main effect on its overall quality ($F(3, 31) = 3.324; p < 0.05$). With the more precise information in the decision-focused summary display, the participants are able to generate decision minutes of higher quality.

Although we did not find any significant effect on other measures of report quality, a finer-grained comparison shows that there are differences between the Baseline condition and the other three conditions in terms of conciseness and trustworthiness. The general purpose display performs fundamentally worse than the decision-focused summary on these two aspects.

<table>
<thead>
<tr>
<th>Criterion (1-7)</th>
<th>TOP</th>
<th>AD-REF</th>
<th>AD-ASR</th>
<th>BASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Quality</td>
<td>2.5</td>
<td>2.4</td>
<td>3.6</td>
<td>3.9</td>
</tr>
<tr>
<td>Completeness</td>
<td>3.1</td>
<td>2.9</td>
<td>3.8</td>
<td>3.4</td>
</tr>
<tr>
<td>Conciseness</td>
<td>2.4</td>
<td>2.7</td>
<td>2.6</td>
<td>3.4</td>
</tr>
<tr>
<td>Writing Style</td>
<td>2.6</td>
<td>2.1</td>
<td>3.3</td>
<td>3.4</td>
</tr>
<tr>
<td>Trustworthiness</td>
<td>1.9</td>
<td>2.0</td>
<td>1.8</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Table 4. Quality assessment of the participants’ minutes on a 7-point scale: the lower the score, the better the minute quality.

**Perceived success analysis**

The average ratings reported in the post-questionnaires (cf. Table 5) suggest that the decision-focused display is perceived to be easier to use ($F(3, 31) = 4.819; p < 0.05$) and less demanding in the amount of effort required ($F(3, 31) = 4.343; p < 0.05$). The participants using the decision-focused display also find themselves able to retrieve the relevant information more efficiently ($F(3, 31) =
User Behavior | Measures
--- | ---
Task completeness | Number of meetings that have been read
First decision written | Time to type first character
Proportion of clicked extracts | Number of content clicks, normalized by number of content buttons
Writing speed | Number of content clicks, normalized by length of experiment
Reading speed (Extract) | Average time-stamp of insertions, normalized by experiment length
Productivity (by writing time-stamp) | Number of words in user’s report (edited)
Productivity (by report length) | Number of times user played the audio or video
Usage of media | Number of content clicks in the preceding 2 minutes of a writing tab click
Usage of extracts to correct writing | Table 3. Log file-based measures of task effectiveness.

<table>
<thead>
<tr>
<th>Criterion (1-5)</th>
<th>TOPLINE</th>
<th>AD-REF</th>
<th>AD-ASR</th>
<th>BASELINE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived ease of use (interface)</td>
<td>4.4</td>
<td>4.1</td>
<td>4.3</td>
<td>3.6</td>
</tr>
<tr>
<td>Perceived efficiency</td>
<td>3.9</td>
<td>3.4</td>
<td>3.6</td>
<td>3.3</td>
</tr>
<tr>
<td>Perceived comprehension (general)</td>
<td>4.6</td>
<td>4.6</td>
<td>4.1</td>
<td>4.1</td>
</tr>
<tr>
<td>Perceived task success (decision)</td>
<td>4.3</td>
<td>4.3</td>
<td>3.8</td>
<td>3.7</td>
</tr>
<tr>
<td>Perceived task difficulty</td>
<td>2.6</td>
<td>2.9</td>
<td>2.9</td>
<td>3.7</td>
</tr>
<tr>
<td>Perceived pressure</td>
<td>2.8</td>
<td>3.8</td>
<td>2.7</td>
<td>3.4</td>
</tr>
<tr>
<td>Perceived system usefulness</td>
<td>4.4</td>
<td>4.3</td>
<td>4.1</td>
<td>4.1</td>
</tr>
</tbody>
</table>

Table 5. User perceived task success. Results are obtained on a 5-point scale (5 = agree strongly, and 1 = disagree strongly).

8.710; \( p < 0.01 \), and absorb the decisions made in the meetings more effectively \( (F(3, 31) = 4.714; p < 0.05) \).

User behavior analysis

An ANOVA test on the log data reveals that, compared to the participants in the baseline condition (automatic general-purpose extract, ASR transcript), participants use a significantly higher proportion of the extracted decision DAs to write minutes \( (F(3, 31) = 9.878; p < 0.001) \) and rely more on the extract contents to modify their minutes \( (F(3, 31) = 21.715; p < 0.001) \). (See Table 6.)

Decision-focused v.s. General-purpose Extracts

Having examined the main effects across the four conditions, we now investigate whether there exists significant differences between each pair of conditions. We then perform Tukey HSD tests to conduct multiple comparisons across each pair of conditions. In particular, we focus on three pairs of conditions: AE v.s. AD, AD v.s. MD, and ASR v.s. REF.

The first pair we compare is the AD and AE conditions. Given that the decision-focused summaries are more effective than the general-purpose summaries for the decision debriefing task, we wish to determine whether the effectiveness remains when a more error-prone automatically generated summary is used in the interface. To determine patterns that were not specified a priori, a posteriori pairwise comparisons were performed.

First, we examined the decision hits across the condition that uses the automatic general summary (BASELINE) and that uses the automatic decision-focused one (AD-ASR). The percentage of decision hits (in Fig. 3) shows that focusing on only the decision-related information results in greater task effectiveness—on average, increasing the number of decision hits over that yielded with the general-purpose display by 36%. Moreover, the decision minutes generated by the participants who used the decision-focused summaries tend to exhibit better overall quality and conciseness.

Further analysis of user behaviors reveal that the participants still rely more on the decision-focused display to summarize meeting decisions, even when the summary contains ASR errors. The decision-focused display is found to significantly increase the use of the summary display \( (p < 0.001) \), normalized frequency of switching to the writing tab \( (p < 0.05) \), and the usage of the summary display prior to writing correction \( (p < 0.001) \).

Automatic Extracts v.s. Manual Extracts

The second pair we compare is the AD and MD conditions. The question that emerges naturally next is how much performance degradation results from replacing the manual summary with its automatic version (which contains 30%-40% inconsistencies with the ground truth). The answer may provide useful guidance for the design of meeting browsers, and may provide support for the development of automatic machinery for query-focused speech summarization.

To answer this question, we compared task effectiveness and report quality of the condition that uses manual decision-focused summaries (TOPLINE) to the one that uses automatically generated decision-focused extracts (AD-REF). Although the participants produced decisions minutes of good quality in all conditions, participants who were presented with the automatic extractive summaries (AE) on average yielded three fewer decision hits (21%).

To further understand whether the errors in the summary...
resulted in any systematic difference in user behavior, we examine the log files (cf. Table 6). We expected that users would prefer to use the meeting summary display to find decisions when the summaries are reflective of the actual decisions made. The post-hoc test results match the expectation: Using the automatic version of the summary (Column AD-REF) instead of the manual version (Column TOPLINE) significantly decreased the use of the summary display prior to writing correction ($p < 0.01$).

However, this difference in task effectiveness and user behavior does not seem to affect the participants’ perceived success towards the task and ability to produce quality minutes: no significant difference was found in any of the participants’ ratings in the post-questionnaire or the report quality ratings for the two conditions.

**Effect of Transcription Type**

The third pair we compare is the ASR and REF conditions. Because our ultimate goal is to design a meeting browser that can be used as soon as a meeting ends, it is important to study whether operating the browser on error-prone ASR transcription (with 30%-40% errors) affects task effectiveness and report quality.

To examine the performance degradation caused by the ASR transcript display, post-hoc tests were also performed across the conditions that operate on ASR transcription (AD-ASR) and on manual transcription (AD-REF). The assessment results of report quality (cf. Bars AD-REF and AD-ASR in Fig. 3) suggest that displaying decision-focused summaries on manual transcripts helps the participants find 39% more (on average, 4 to 5) decision hits than displaying the summaries based on ASR transcripts ($p < 0.01$, Tukey’s test).

Further analysis of the decision minute quality (cf. Table 4) shows that users who browse summaries based on manual transcripts are likely to produce decision minutes of better overall quality and completeness. In addition, the more readable transcripts allow the participants to allocate more of their time to absorbing relevant information, rather than understanding meeting content. Thus, the decision minutes generated by this group of users may be better appreciated by readers.

Examination of user behaviors (cf. Column AD-REF and AD-ASR in Table 6) also shows the transcription type to have effects on the usage of the summary display for writing decision minutes ($p < 0.05$). The less helpful displays also increase the level of perceived pressure ($p < 0.05$) reported by the participants (cf. Row 6 in Table 5). Compared to the errors introduced by the automatic decision-focused extracts, the errors introduced by the ASR transcripts have a greater negative impact on the users’ performance level.

**DISCUSSION**

The results of this study verify our experimental hypothesis. Displaying decision-focused summaries in the meeting browser helps users to obtain an overview of the decisions from multiple meeting recordings more effectively than general-purpose summaries. It also increases users’ perceived efficiency. Decision-focused summaries, obtained by filtering out the dialogue acts irrelevant to decisions, were found to improve not only task effectiveness, but also the overall quality of the participants’ minutes. The users in the focused summary conditions read through a higher proportion of summary material to find relevant information and relied more on summaries to prepare and correct the decision minutes they wrote.

Our investigation further examined the impact of using automatically generated decision-focused summaries and operating a meeting browser on ASR transcripts. The first examination showed that, even when the displayed decision summaries were automatically generated, participants who used the decision-focused summary display still outperformed those who used the general-purpose summary display in the decision debriefing task. Although the automatic summary users did not achieve the same level of task effectiveness as those using manual summaries, they were able to produce decision minutes of similar quality.

One explanation for this could be that parts of the automatic summary correctly identified some of the decision points, and users leveraged the correct parts to find information relevant to these decision points for summarization. In fact, the user behavior analysis found task effectiveness (i.e., the number of decisions hits) and usage of the summary display (i.e., the proportion of the summary that was used to prepare and correct writing) to be significantly correlated (Spearman’s test; $r = 3.573, p < 0.001$).

**Use of audio-video aids**

From the post-experiment debriefings, we observed two main strategies adopted by users to find the decision points that were not clearly presented in the summaries: (1) Some users attempted to go through the extracts in the summary display one by one, looking for relevant information in the surrounding context in the transcript; (2) Others turned to the audio-video recordings to find the missing decisions. The two coping strategies can be distinguished by their usage of media. Table 7 presents the proportion of par-

<table>
<thead>
<tr>
<th>Proportion of clicked extracts</th>
<th>TOPLINE</th>
<th>AD-REF</th>
<th>AD-ASR</th>
<th>BASELINE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage of media</td>
<td>23.60</td>
<td>17.71</td>
<td>33.44</td>
<td>15.56</td>
</tr>
<tr>
<td>Usage of extracts to correct writing</td>
<td>6.84</td>
<td>1.54</td>
<td>0.93</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 6. Task effectiveness measures based on user behavioral cues.
Participants that have high and low usage of the audio video aids.

It appears that when the manual transcripts are in the display, e.g., in the Topline and AD-REF conditions, the choice of strategy was based on individual differences, and a majority of the users preferred to use the decision-focused summaries rather than the audio-video aids. Yet when the error-prone ASR transcripts are in the display, e.g., in the AD-ASR and Baseline conditions, the choice of strategy was noticeably affected by the type of summary display. Comparing Columns AD-ASR and Baseline in Table 7 illustrates that the AD-ASR users tended to make more usage of the audio-video recordings. This is because the ASR transcripts are difficult to understand by themselves, and it is therefore important to find additional hints from the summary display. However, as the summaries presented in the AD-ASR display are often short and error-prone, the audio-video recordings are necessary for accomplishing the task.

<table>
<thead>
<tr>
<th>Media Usage</th>
<th>Topline</th>
<th>AD-REF</th>
<th>AD-ASR</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low (&lt; 30)</td>
<td>70.0%</td>
<td>85.7%</td>
<td>44.4%</td>
<td>88.9%</td>
</tr>
<tr>
<td>High (≥ 30)</td>
<td>30.0%</td>
<td>14.3%</td>
<td>55.6%</td>
<td>11.1%</td>
</tr>
</tbody>
</table>

Table 7. The proportion of participants who had low and high usage of audio-video recordings: Low=playing recordings less than 30 times; High=playing recordings greater than or equal to 30 times.

Fig. 4 demonstrates the effect of audio-video usage on task effectiveness and perceived success. The analysis reveals that the AD-ASR and Baseline users who turned to the audio-video browsing strategy (i.e., those with high usage of audio-video aids) were more likely to miss decisions. Interestingly, the lower task success rates did not affect the ratings of user-perceived success. For example, the high media usage participants under the AD-ASR condition, who tended to be less effective in decision debriefing, still perceived a high level of task success. The finding coincides with the participants’ comments that, although the audio-video recordings are difficult to use, they provided grounds for decision understanding.

CONCLUSIONS

This study has verified our experimental hypothesis: Existing meeting summarization systems, which provide a general-purpose summary display (which represents compressions of 30-40%), can be improved by refocusing the summaries with regard to the user’s information need (which represent compressions of 1-2%). For users who require a quick overview of decisions, the decision-focused summary display was found to improve not only the task effectiveness, but also the overall report quality.

Users also perceived the decision-focused summaries useful in helping them to find all relevant information and understand the decisions more efficiently. However, as the decision debriefing task in our experimental design is considered to be a straightforward task to complete within the time allowed, we cannot draw conclusions about the impact of the decision-focused extracts on efficiency. The browser interfaces that come with the decision-focused summary display are also rated as easier to use.

In addition, we evaluated the impacts of automation on the decision debriefing task. The findings are as follows: (1) The automatically generated decision-focused summaries, which contain 30%-40% inconsistencies with the gold standard manual summaries (as an average annotator), still assist users in producing high quality decision minutes and feeling confident about their performance on the decision debriefing task, despite some reduction in decision hits. (2) The ASR transcription has a greater negative impact on the actual task effectiveness and the quality of minutes. Another side effect of the ASR display is an increase in the level of user-perceived pressure.

Further investigation demonstrates a correlation between task effectiveness and usage of the summary display. As the content in the decision-focused summary is more closely tied to user needs, participants who use these summaries (as opposed to the general purpose ones) rely more on the summary to find relevant information and, in turn, achieve higher performance.

Finally, the examination of users’ media usage and coping strategies suggest that there exists an individual difference
in the user’s preference of whether to use the summary display or the audio-video playback facility to find relevant information. However, when the decision-focused summary is displayed with the ASR transcripts, users are often forced to view the audio-video recordings, since it is too difficult to use the other two displays. This hindered the performance of users who do not prefer to play-back the audio-video recordings. This also suggests that there is a need to provide additional interface assistance to facilitate this group of users when ASR transcripts are used and may affect the comprehensibility of a succinct decision-focused summary. Possible interface enhancements include a switching device that allows users to freely go from the view of a decision-focused summary back to that of a general-purpose summary.

In summary, this evaluation provides the first account of how query-focused summaries facilitate users’ understanding of multimedia archives. However, there are concerns about whether our approach is successful only when used to summarize well-structured meetings that follow a scenario like the ones in the AMI corpus, where meetings typically end with a summary. These concerns have first been addressed by designing the meeting scenario used in the AMI corpus collection to be as close to real-life meetings as possible. In addition, the meeting participants do not always summarize decisions at the end. Empirical analysis of the positions of the decision DAs confirmed this: only 26% of the decision DAs occur in the last five minutes of the meeting, and the average position of the decision DAs is in the middle portion of the meetings (around 60% into the meetings). Last but not least, since the participants in this evaluation were not provided with any material relevant to the meeting scenario other than the material available through the meeting browser, this evaluation is expected to reflect how users will employ the focused summary-enhanced browser to gather information needed for debriefing query-related information.

Acknowledgments
This work is supported by the AMIDA project (European IST Programme Project FP6-0033812). This paper only reflects the authors’ views and funding agencies are not liable for any use that may be made of the information contained herein. We thank Gabriel Murray, Theresa Wilson, and Jean Carletta for their valuable feedbacks and Jonathan Kilgour for his continuous engineering support. We also thank the anonymous reviewers for their valuable feedbacks.

REFERENCES
17. V. Pallotta, J. Niekrasz, and M. Purver. Collaborative and argumentative models of meeting discussions. In Workshop on Computational Models of Natural Arguments (CMNA) at the IJCAI, 2005.