Extrinsic Summarization Evaluation: A Decision Audit Task

Gabriel Murray[1], Thomas Kleinbauer[2], Peter Poller[2], Steve Renals[3], Jonathan Kilgour[3], and Tilman Becker[2]

[1] University of British Columbia, Vancouver, Canada
[2] German Research Center for Artificial Intelligence, Saarbrücken, Germany
[3] University of Edinburgh, Edinburgh, Scotland

Abstract. In this work we describe a large-scale extrinsic evaluation of automatic speech summarization technologies for meeting speech. The particular task is a decision audit, wherein a user must satisfy a complex information need, navigating several meetings in order to gain an understanding of how and why a given decision was made. We compare the usefulness of extractive and abstractive technologies in satisfying this information need, and assess the impact of automatic speech recognition (ASR) errors on user performance. We employ several evaluation methods for participant performance, including post-questionnaire data, human subjective and objective judgments, and an analysis of participant browsing behaviour.

1 Introduction

In the field of automatic summarization, machine summaries are often evaluated *intrinsically*, i. e., according to how well their information content matches the information content of multiple reference summaries. A more comprehensive and reliable evaluation of the quality of a given summary, however, is the degree to which it aids a real-world *extrinsic* task: an indication not just of how informative the summary is, but how useful it is in addressing a real information need. While intrinsic evaluation metrics are indispensable for development purposes and can be easily replicated, they ideally need to be chosen based on whether or not they are good predictors for extrinsic usefulness, e.g. whether they correlate to a measure of real-world usefulness.

We therefore design an extrinsic task that models a real-world information need, create multiple experimental conditions and enlist subjects to participate in the task. The chosen task is a *decision audit*, wherein a user must review previously held meetings in order to determine how a given decision was reached. This involves the user determining what the final decision was, which alternatives had previously been proposed, and what the arguments for and against the various proposals were. The reason this task was chosen is that it represents one of the key applications for analyzing multimodal interactions - that of aiding *corporate memory*, the storage and management of a organization's knowledge, transactions, decisions, and plans. A organization may find itself in the position

A. Popescu-Belis and R. Stiefelhagen (Eds.): MLMI 2008, LNCS 5237, pp. 349–361, 2008.

of needing to review or explain how it came to a particular position or why it took a certain course of action. We hypothesize that this task will be made much more efficient when meetings are archived and summarized.

The decision audit represents a complex information need that cannot be satisfied with a simple one-sentence answer. Relevant information will be spread throughout several meetings and may appear at multiple points in a single discussion thread. Because the decision audit does not only involve knowing *what* decision was made but also determining *why* the decision was made, the person conducting the audit will need to understand the evolution of the meeting participants' thinking and the range of factors that led to the ultimate decision. Because the person conducting the decision audit does not know which meetings are relevant to the given topic, there is an inherent relevance assessment task built into this overall task. As time is limited, they cannot hope to scan the meetings in their entirety and so must focus on which meetings and meeting sections seem most promising.

2 Related Extrinsic Evaluation Work

This section describes previous extrinsic evaluations relating either to summarization or to the browsing of multi-party interactions. We then describe how our decision audit browsers fit into a typology of multi-media interfaces.

2.1 Previous Work

In the field of text summarization, a commonly used extrinsic evaluation has been the *relevance assessment* task [1]. In such a task, a user is presented with a description of a topic or event and then must decide whether a given document (e.g. a summary or a full-text) is relevant to that topic or event. Such schemes have been used for a number of years and on a variety of projects [2, 3, 4]. Due to problems of low inter-annotator agreement on such ratings, Dorr et. al [5] proposed a new evaluation scheme that compares the relevance judgment of an annotator given a full text with that same annotator given a condensed text.

Another type of extrinsic evaluation for summarization is the *reading comprehension* task [1, 6, 7]. In such an evaluation, a user is given either a full source or a summary text and is then given a multiple-choice test relating to the full source information. A system can then calculate how well they perform on the test given the condition. This evaluation framework relies on the idea that truly informative summaries should be able to act as substitutes for the full source.

In the speech domain, there have been several large extrinsic IR evaluations in the past few years, though not necessarily designed with summarization in mind. Wellner et. al [8] introduced the Browser Evaluation Test (BET), in which *observations of interest* are collected for each meeting, e.g. the observation "Susan says the footstool is expensive." Each observation is presented as both a positive and negative statement and the user must decide which statement is correct by browsing the meetings and finding the correct answer. It is clear that such a setup could be used to evaluate summaries and to compare summaries with other information sources. We choose not to use this evaluation paradigm, however, because the observations of interest tend to be skewed towards a keyword search approach, where it would always be simpler just to search for a word such as "footstool" rather than read a summary.

The Task-Based Evaluation (TBE) [9] evaluates multiple browser conditions containing various information sources relating to a series of meetings. Participants are brought in four at a time and are told that they are replacing a previous group and must finish that group's work. In essence, the evaluation involves re-running the final meetings of the series with new participants. The participants are given information related to the previous group's initial meetings and must finalize the previous group's decisions as best as possible given what they know. There are several reasons we have chosen not to use the TBE for this summarization evaluation. One is that the TBE relies primarily on post-questionnaire answers for evaluation. While we do incorporate post-questionnaires in our evaluation, we are also very interested in the objective participant performance in the task and browsing behaviour during the task. Two, the TBE is more costly to run than our decision audit task, as it requires having groups of four people spend an afternoon reviewing previous meetings and conducting their own meetings, which are also recorded, whereas the decision audit is an individual task.

The SCANMail browser [10, 11] is an interface for managing and browsing voicemail messages, with multi-media components such as audio, ASR transcripts, audio-based paragraphs, and extracted names and phone numbers. To evaluate the browser and its components, the authors compared the SCANMail browser to a state-of-the-art voicemail system on four key tasks: scanning and searching messages, extracting information from messages, tracking the status of messages (e.g. whether or not a message has been dealt with), and archiving messages. Both in a think-aloud laboratory study and a larger field study, users found the SCANMail system outperformed the comparison system for these extrinsic tasks. The field study in particular yielded several interesting findings. In 24% of the times that users viewed a voicemail transcript with the SCANMail system, they did not resort to playing the audio. This testifies to the fact that the transcript and extracted information can, to some degree, act as substitutes for the signal, which user comments also back up. On occasions when users did play the audio, 57% of the time they did not play the entire audio. Most interestingly, 57% of the audio play operations resulted from clicking within the transcript. The study also found that users were able to understand the transcripts even with recognition errors, partly by having prior context for many of the messages.

Whittaker et. al [12] described a task-oriented evaluation of a browser for navigating meeting interactions. The browser contains a manual transcript, a visualization of speaker activity, audio and video streams with play, pause and stop commands, and artefacts such as slides and whiteboard events (the slides, but not the whiteboard events, are indices into the meeting record). Users were given two sets of questions to answer, the first set consisting of general "gist" question about the meeting, and the second set comprised of questions about specific facts within the meeting. There were 10 questions in total to be answered. User responses were subsequently scored on correctness compared with model answers. While general performance was not high, users found it much easier to answer specific questions than "gist" questions using this browser setup. This has special relevance for our work, as certain types of information needs might be easily satisfied without recourse to derived data such as summaries or topic segments, but getting the general gist of the meeting seems to be much more difficult. Very interestingly, users often felt that they had performed much better than they actually had. Specifically, users seemed to be unaware that they had missed relevant or vital information and felt that they had provided comprehensive answers. Across the board, participants focused on reading the transcript rather than beginning with the audio and video records directly.

2.2 Multimodal Browser Types

Tucker and Whittaker [13] provided an overview of the mechanisms available for browsing multimodal meetings. They established a four-way browser classification: audio-based browsers, video-based browsers, artefact-based browsers, and derived data browsers. In light of this classification scheme, our decision audit browsers are video browsers incorporating derived data forms. Although other incarnations of our browsers contain meeting artefacts such as slides, we simplify the browsers as much as possible for this task by putting the focus on derived data forms and their usefulness for browsing the meeting records. Each version of the experimental browser is built using JFerret [14], an easily modifiable multi-media browser framework[1].

3 Task Overview

The experiment consists of five different conditions, described below. We recruited 10 subjects per condition for a total of 50 subjects, all native speakers of English. For each condition, 6 participants were run in Edinburgh and 4 were run at Saarbrücken, the experimental setups for the two locations being as identical as possible.

As our underlying data we chose four meetings from the AMI Meeting Corpus [15]. The meeting series ES2008 was selected because the participant group in that series worked well together on the task of designing a new remote control. The group took the task seriously and exhibited deliberate and careful decision-making processes in each meeting and across the meeting series as a whole.

The basic task for the participants was to write a summary of the decision making process in the meetings for separating often and rarely used functions of the remote control. This particular information need was chosen because the relevant discussion manifested itself throughout the 4 meetings, and the group went through several possibilities before designing an eventual solution to this portion of the design problem. A participant in the decision audit task therefore

[1] http://www.idiap.ch/mmm/tools/jferret

would have to consult each meeting to be able to retrieve the full answer to the task's information need.

Each participant in our task was first given general instructions explaining the meeting browser used in the experiment, the specific information need they were meant to satisfy in the task, and a notice of the allotted time, 45 minutes, which included both searching for the information and writing up the answer. This amount of time was based on the result of an individual pilot task for Condition EAM (s. 3.1). After reading the task instructions, each participant is briefly shown how to use the browser's various functions for navigating and writing in the given experimental condition. They are then given several minutes to familiarize themselves with the browser using unrelated meeting data, until they state that they were comfortable and ready to proceed.

3.1 Experimental Conditions

There are five conditions run in total: one baseline condition, two extractive conditions and two abstractive conditions, all of which come with audio/video recordings and either a manual or automatic meeting transcript. Table 1 lists the experimental conditions. The three-letter ID for each condition corresponds to **k**eywords/**e**xtracts/**a**bstracts, **a**utomatic/**s**emi-automatic/**m**anual algorithms, and **a**utomatic/**m**anual transcripts.

Table 1. Experimental Conditions

Condition	Description
KAM	Top 20 keywords
EAM	Extractive summary of manual transcripts
EAA	Extractive summary of ASR transcripts
AMM	Human abstracts
ASM	Semi-Automatic abstracts

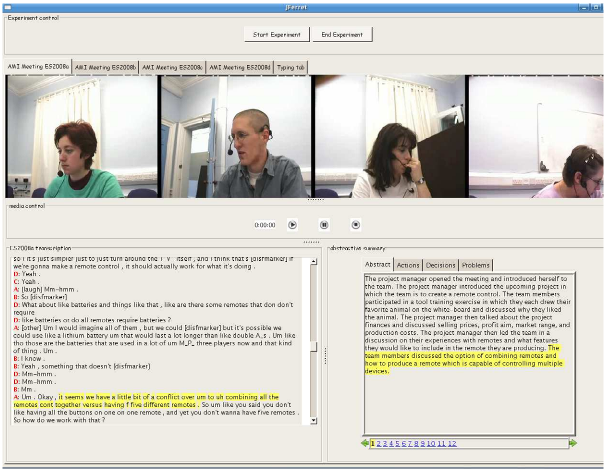
The baseline condition, Condition KAM, consists of a browser with manual transcripts and a list of the top 20 keywords in the meeting. The keywords are determined automatically using *su.idf* [16]. Though this is a baseline condition, the fact that it utilizes manual transcripts gives users in this condition a possible advantage over users in conditions with ASR. In this respect, it is a challenging baseline. There are other possibilities for the baseline, but we choose the top 20 keywords because we are interested in comparing different forms of derived content from meetings, and because a facility such as keyword search would likely be problematic for a participant who is uncertain of what to search for because they are unfamiliar with the meetings.

Condition AMM is the gold-standard condition, a human-authored abstractive summary. Each summary is divided into subsections: abstract, actions, decisions and problems. Because of the distinct "decisions" subsection, this is considered a challenging gold-standard to match for a decision audit task.

Conditions EAM and EAA present the user with an extractive summary of each meeting, with the difference between the conditions being that the latter is based on ASR and the former on manual transcripts. Condition EAA is the only experimental condition using ASR output. These summaries were generated by training a support vector machine (SVM) with an RBF kernel on the AMI training data, using 17 features from five broad feature classes: prosodic, lexical, length, structural and speaker-related. The classifier was run on the four meetings of interest, ranking dialogue acts in descending order of informativeness according to posterior probability, extracting until we reach the desired summary length, approximately 1000 words for the first meeting, 1900 words each for the second and third meetings, and 2300 words for the final meeting. These lengths correlate to the lengths of the meetings themselves and represent compressions to approximately 40%, 32%, 32% and 30% of the total meeting word counts, respectively. These summary lengths were based on the compression rates of the human extracts for these meetings.

Condition ASM presents the user with a semi-automatically generated abstractive summary, as described in [17]. This method utilizes hand-annotated topic segmentation and topic labels available in the AMI corpus. In addition, the meeting transcript was manually annotated with content items from a taxonomy for the domains *project*, *meeting* and *product*. A sentence is generated for each meeting topic based on the annotated topic label. It may also mention the three most frequent content items, indicating roughly what was discussed.

3.2 Browser Setup



so I it's just simpler just to just turn around the T_V_ itself , and I think that's [disfmarker] if we're gonna make a remote control , it should actually work for what it's doing .
D: Yeah .
C: Yeah .
A: [laugh] Mm-hmm .
B: So [disfmarker]
D: What about like batteries and things like that , like are there some remotes that don don't require
D: like batteries or do all remotes require batteries ?
A: [other] Um I would imagine all of them , but we could [disfmarker] but it's possible we could use like a lithium battery um that would last a lot longer than like double A_s . Um like tho those are the batteries that are used in a lot of um M_P_ three players now and that kind of thing . Um .
B: I know .
B: Yeah , something that doesn't [disfmarker]
D: Mm-hmm .
D: Mm-hmm .
B: Mm .
A: Um . Okay , it seems we have a little bit of a conflict over um to uh combining all the remotes cont together versus having f five different remotes . So um like you said you don't like having all the buttons on one on one remote , and yet you don't wanna have five remotes . So how do we work with that ?

The project manager opened the meeting and introduced herself to the team. The project manager introduced the upcoming project in which the team is to create a remote control. The team members participated in a tool training exercise in which they each drew their favorite animal on the white-board and discussed why they liked the animal. The project manager then talked about the project finances and discussed selling prices, profit aim, market range, and production costs. The project manager then led the team in a discussion on their experiences with remotes and what features they would like to include in the remote they are producing. The team members discussed the option of combining remotes and how to produce a remote which is capable of controlling multiple devices.

Fig. 1. Condition AMM Browser

The meeting browsers are kept essentially the same in all conditions to eliminate any potential confounding factors relating to the user interface. In each browser, there are 5 tabs for the 4 meetings and a writing pad, provided for the participant to author their decision audit answer. As a consequence, the participant cannot view the meeting tabs while typing the answer; they are restricted to tabbing back and forth as needed. This was designed deliberately so as to be able to discern when the participant was working on formulating or writing the answer on the one hand and when they were browsing the meeting records on the other. In each meeting tab, the videos displaying the four meeting participants are laid out horizontally with the media controls beneath. The transcript is shown in the lower left of the browser tab in a scroll window.

In Condition KAM, each meeting tab contains buttons corresponding to the top 20 keywords. Pressing a button highlights the first instance of the associated keyword in the transcript, as well as opening a list of hyperlinks to all occurrences of that word in the transcript. In Conditions AMM and ASM, the abstractive summary is presented next to the meeting transcript. Clicking on a summary sentence opens a list of hyperlinks similar to Condition KAM, linking to dialogue acts in the transcript that support the particular summary sentence. In addition to an abstract, Condition ASM displays three extra tabs with bullet points for the subsections mentioned above (s. fig. 1). In Conditions EAM and EAA, the extractive summary is displayed with each dialogue act hyperlinked to the point in the transcript it was extracted from.

3.3 Evaluation Features

For evaluation of the decision audit task, there are three types of features to be analyzed: answers to questionnaires, human ratings of the users' written answers, and features extracted from logfiles. In all conditions, we log with time stamps mouse clicks on the transcript, the play-, pause-, stop buttons, changing tabs, and characters entered into the typing tab.

Upon completion of the decision audit task, we present each participant with a post-task questionnaire consisting of 10 statements with which the participant can state their level of agreement or disagreement via a 5-point Likert scale, such as *I was able to efficiently find the relevant information*, and two open-ended questions about the specific type of information available in the given condition and what further information they would have liked. Of the 10 statements evaluated, some are re-wordings of others with the polarity reversed in order to gauge the users' consistency in answering.

In order to gauge the participant accomplished the decision audit task, we enlist two human judges to do both *subjective* and *objective* evaluations. For the subjective portion, the judges first read through all 50 answers to get a view of the variety of answers. They then rate each answer using a 1-8 Likert-scale on criteria roughly relating to the precision, recall and f-score of the answer, as well as effort, comprehension and writing style (s. table 3). The results are averaged to yield a single score. For the objective evaluation, three judges constructed a gold-standard list of 25 items that should be contained in an ideal answer to the decision audit task. Two of them then checked off individually how many of the gold-standard items were contained in each participant answer. In a second step, they identified those participant answers where their ratings diverged by more than two points. There were 12 out of 50 ratings pairs that needed revision in this manner. After the judges' consultation on those 12 pairs of ratings, each experiment was given a single objective rating.

4 Results

Post-Questionnaires. An analysis of the *post-questionnaires* reveals that participants in general find the task to be challenging, as evidenced by the average answers on questions 4, 6 and 7 in Table 2. The task was designed to be challenging and time-constrained, because a simple task with a plentiful amount of

Table 2. Post-Questionnaire Results

Question	KAM	EAM	EAA	AMM	ASM
Q1: *I found the meeting browser intuitive and easy to use*	3.8	4.0	3.02_{AMM}	$4.3^{EAA,ASM}$	3.7_{AMM}
Q2: *I was able to find all of the information I needed*	2.9_{AMM}	3.8	2.9_{AMM}	$4.1^{KAM,EAA,ASM}$	3.0_{AMM}
Q3: *I was able to efficiently find the relevant information*	2.8_{AMM}	3.4^{ASM}	2.5_{AMM}	$4.0^{KAM,EAA,ASM}$	$2.65_{EAM,AMM}$
Q4: *I feel that I completed the task in its entirety*	2.3_{AMM}	3.1	2.3	3.2^{KAM}	2.9
Q5: *I understood the overall content of the meeting discussion*	3.8	**4.5**	3.9	4.1	3.9
Q6: *The task required a great deal of effort*	3.0	2.6^{EAA}	3.9_{EAM}	3.1	3.2
Q7: *I had to work under pressure*	3.3	**2.6**	3.3	2.7	3.1
Q8: *I had the tools necessary to complete the task efficiently*	3.1_{EAM}	$4.3^{KAM,EAA,ASM}$	3.0_{EAM}	4.1	3.5_{EAM}
Q9: *I would have liked additional information about the meetings*	3.0_{EAM}	2.0^{KAM}	2.4	2.6	2.7
Q10: *It was difficult to understand the content of the meetings...*	2.1	$1.5^{EAA,ASM}$	2.7_{EAM}	2.0	2.3_{EAM}

For each score in the table, that score is significantly better than the score for any conditions in superscript, and significantly worse than the score for any condition in subscript (according to t-test).

allotted time would allow the participants to simply read through the entire transcript or listen and watch the entire audio/video record in order to retrieve the correct information, disregarding other information sources. The task as designed requires efficient navigation of the information in the meetings in order to finish the task completely and on time.

Participants in condition AMM found the gold-standard human abstracts and specifically the summary subsections to be very valuable sources of information. One participant remarked “Very well prepared summaries. They were adequate to learn the jist [sic] of the meetings by quickly skimming through... I especially liked the tabs (Decisions, Actions, etc.) that categorised information according to what I was looking for.”

Condition ASM rated quite well on questions regarding ease of use and intuitiveness, but slightly less well in terms of using the browser to locate the important information. It does consistently rate better than KAM and EAA.

For overall comprehension of the information in the meetings, extractive summaries were rated the highest of all. Extractive summaries of manual transcripts (EAM) were also rated the best in terms of the effort required to conduct the task. Perhaps the most compelling result is that Condition EAM not only rated the best in a question relating to having the tools necessary to complete the task, but it is *significantly better* than all conditions except the gold-standard human abstracts (according to t-test).

However, it is quite clear that the errors within an ASR transcript adversely affect user satisfaction in such an information retrieval task. For the questions relating to the effort required, the tools available, and the difficulty in understanding the meetings, Condition EAA tends to perform the worst of all, on par or even lower than the baseline condition. It should be noted however, that a baseline such as Condition KAM is working off of *manual* transcripts and would be expected to be worse when applied to ASR. As mentioned earlier, the baseline is a challenging baseline in that respect. Judging from the open-ended questions

Table 3. Human Evaluation Results - Subjective and Objective

Criterion	**KAM**	**EAM**	**EAA**	**AMM**	**ASM**
Q1: *overall quality*	3.0_{AMM}	4.15	3.05_{AMM}	$4.65^{KAM,EAA}$	4.3
Q2: *conciseness*	$2.85_{EAM,AMM,ASM}$	4.25^{KAM}	3.05_{AMM}	$4.85^{KAM,EAA}$	4.45^{KAM}
Q3: *completeness*	2.55_{AMM}	3.6	2.6_{AMM}	$4.45^{KAM,EAA}$	3.9
Q4: *task comprehension*	$3.25_{EAM,AMM}$	$\mathbf{5.2}^{KAM,EAA}$	$3.65_{EAM,AMM}$	$5.25^{KAM,EAA}$	4.7
Q5: *participant effort*	4.4	$\mathbf{5.2}^{EAA}$	$3.7_{EAM,AMM,ASM}$	5.3^{EAA}	4.9^{EAA}
Q6: *writing style*	4.75	5.65^{EAA}	$4.1_{EAM,AMM,ASM}$	5.7^{EAA}	5.8^{EAA}
Q7: *objective rating*	4.25_{AMM}	7.2	5.05_{AMM}	$9.45^{KAM,EAA}$	7.4

For each score in the table, that score is significantly better than the score for any conditions in superscript, and significantly worse than the score for any condition in subscript (according to t-test).

in the post-questionnaires, it's clear that at least two participants found the ASR so difficult to work with that they tended not to use the extractive summaries, let alone the full transcript, relying instead on watching the audio/video as much as possible.

Subjective Evaluation. Table 3 shows the results of the *subjective* evaluation. Condition AMM is clearly a challenging gold-standard, and Conditions EAM and ASM are roughly comparable to each other. Subjective ratings drop off sharply for Condition EAA incorporating ASR, particularly for comprehension and writing style. We presume that the ASR errors cause participants in that condition to have a lower understanding of the meeting content, which in turn leads to lower coherence and inferior writing quality in their responses. Interestingly, the scores on each criterion and for *every* condition tend to be somewhat low on the Likert scale, due to the difficulty of the task.

Objective Evaluation. According to the objective evaluation, Condition AMM is superior, with an average more than two points higher than the next best condition. The worst overall is the baseline Condition KAM, averaging only 4.25 hits (of a maximum possible 25). However, while the worst two conditions are significantly worse than the best overall condition, there are no significant differences between the other pairs of conditions, e.g. Condition EAA incorporating ASR is not significantly worse than Conditions EAM and ASM. So even with an errorful transcript, participants in Condition EAA are able to retrieve the relevant pieces of information at a rate not significantly worse than participants with a manual transcript. The quality may be worse from a subjective standpoint, as evidenced in the previous section, but the decision audit answers are still informative and relevant.

For the objective evaluation, in any given condition there is a large amount of variance that is simply down to differences between users. For example, even in the gold-standard Condition AMM there are some people who can only find one or two relevant items whilst others find 16 or 17. Given a challenging task and a limited amount of time, some people may have simply felt overwhelmed in trying to locate the informative portions efficiently.

Browsing Evaluation. A result gleaned from close analysis of participants' browsing behaviours shows an interesting strategy of people in Condition EAA faced

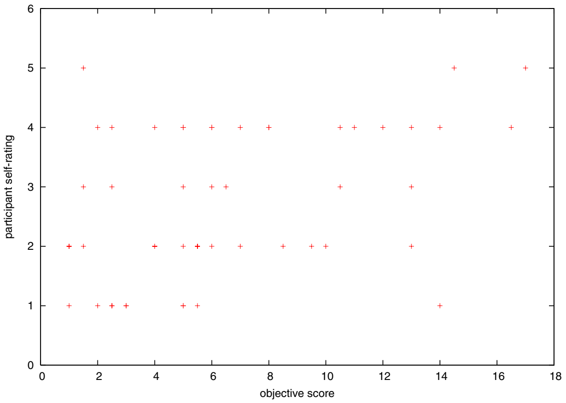


Fig. 2. Objective Scores and Post-Questionnaire Scores

with ASR transcripts. While they still frequently use the summary dialogue acts as indices into the meeting record, they subsequently utilize the audio/video record much more frequently than in the other conditions, presumably to disambiguate the errors encountered. This shows that, to some extent, participants can compensate for the noisy transcript by altering their browsing strategies, using the summaries in tandem with the audio/video in order to find the relevant items from the meetings.

The analysis of browsing behaviour also shows that participants in the gold-standard Condition AMM are able to begin answering the question much earlier in the task, write longer answers overall, and have more time for editing before times expires.

Questionnaire/Objective Evaluation Correlation. Figure 2 shows the relationship between the objective ratings and participant self-ratings for all 50 participants. While the positive correlation is evident, an interesting trend is that while there are relatively few people who score highly on the objective evaluation but score low on the self-ratings, there are a fair number of participants who have a low objective score but rate themselves highly on the post-questionnaire. A challenge with this type of task is that the participant simply may not have a realistic idea of how much relevant information is out there. After retrieving four or five relevant items, they may feel that they've completed the task entirely. This result is similar to the finding by Whittaker et. al [12], mentioned in the discussion of previous work, where participants often feel that they performed better than they really did.

5 Discussion

Although the semi-automatic abstracts got average reviews in the post-questionnaire, both the subjective and objective evaluation rate them second after the gold standard for most ratings, or even better (writing style). For task comprehension and participant effort, they come in third after EAM, however, the difference in rating is not significant. These are encouraging results for further research in automatic abstracting.

Overall the results are also very good news for the extractive summarization paradigm. Users find extractive summaries to be intuitive, easy-to-use and efficient, are able to employ such documents to locate the relevant information in a timely manner according to human evaluations, and users are able to adapt their browsing strategies to cope with ASR errors. While extractive summaries might be far from what people conceptualize as a traditional meeting summary, they are intuitive and useful documents in their own right.

Perhaps the most interesting result from the decision audit overall is regarding the effect of ASR on carrying out such a complex task. While participants using ASR find the browser to be less intuitive and efficient, they nonetheless feel that they understand the meeting discussions and do not desire additional information sources. In a subjective human evaluation, the quality of the answers in Condition EAA suffers according to most of the criteria, including writing style, but the participants are still able to find many of the relevant pieces of information according to the objective human evaluation. We find that users are able to adapt to errorful transcripts by using the summary dialogue acts as navigation and then relying much more on audio/video for disambiguating the conversation in the dialogue act context. Extractive summaries, even with errorful ASR, are useful tools for such a complex task, particularly when incorporated into a multi-media browser framework. There is also the possibility of creating browsing interfaces that minimize the user's direct exposure to the ASR transcript (e.g. audio summaries with limited textual accompaniment).

6 Conclusion

We have presented an extrinsic evaluation paradigm for the automatic summarization of spontaneous speech in the meetings domain: a decision audit task. In each condition of the experiment, users were able to utilize the derived content in order to find and extract information relevant to a specific task need. The largely positive results for the extractive conditions justify continued research on this summarization paradigm. However, the considerable superiority of gold-standard abstracts in many respects also support the view that research should begin to try to bridge the gap between extractive and abstractive summarization.

It is widely accepted in the summarization community that there should be increased reliance on extrinsic measures of summary quality. It is hoped that the decision audit task will be a useful framework for future evaluation work. Intrinsic and extrinsic methods should be used hand-in-hand, with the former as a valuable development tool and predictor of usefulness and the latter as a real-world evaluation of the state-of-the-art.

Acknowledgements. This work is supported by the European IST Programme Project FP6-033812.

References

1. Mani, I.: Summarization evaluation: An overview. In: Proc. of the NTCIR Workshop 2 Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization, Tokyo, Japan, pp. 77–85 (2001)
2. Jing, H., Barzilay, R., McKeown, K., Elhadad, M.: Summarization evaluation methods: Experiments and analysis. In: Proc. of the AAAI Symposium on Intelligent Summarization, Stanford, USA, pp. 60–68 (1998)
3. Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., Sundheim, B.: The TIPSTER SUMMAC text summarization evaluation. In: Proc. of EACL 1999, Bergen, Norway, pp. 77–85 (1999)
4. Harman, D., Over, P.: Document understanding conference 2004. In: Proc. of the DUC 2004, Boston, USA (2004)
5. Dorr, B., Monz, C., President, S., Schwartz, R., Zajic, D.: A methodology for extrinsic evaluation of text summarization: Does ROUGE correlate? In: ACL 2005, MTSE Workshop, Ann Arbor, USA, pp. 1–8 (2005)
6. Hirschman, L., Light, M., Breck, E.: Deep read: A reading comprehension system. In: Proc. of ACL 1999, College Park, MD, USA, pp. 325–332 (1999)
7. Morris, A., Kasper, G., Adams, D.: The effects and limitations of automated text condensing on reading comprehension performance. Information Systems Research 3, 17–35 (1992)
8. Wellner, P., Flynn, M., Tucker, S., Whittaker, S.: A meeting browser evaluation test. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems 2005, pp. 2021–2024. ACM Press, New York (2005)
9. Kraaij, W., Post, W.: Task based evaluation of exploratory search systems. In: Proc. of SIGIR 2006 Workshop, Evaluation Exploratory Search Systems, Seattle, USA, pp. 24–27 (2006)
10. Hirschberg, J., Bacchiani, M., Hindle, D., Isenhour, P., Rosenberg, A., Stark, L., Stead, L., Whittaker, S., Zamchick, G.: SCANMail: Browsing and searching speech data by content. In: Proc. of Interspeech 2001, Aalborg, Denmark, pp. 1299–1302 (2001)
11. Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., Stead, L., Zamchick, G., Rosenberg, A.: Scanmail: a voicemail interface that makes speech browsable, readable and searchable. In: Proc. of the SIGCHI 2002, Minneapolis, Minnesota, pp. 275–282. ACM, New York (2002)
12. Whittaker, S., Tucker, S., Swampillai, K., Laban, R.: Design and evaluation of systems to support interaction capture and retrieval. Personal and Ubiquitous Computing (to appear)
13. Tucker, S., Whittaker, S.: Accessing multimodal meeting data: Systems, problems and possibilities. In: Bengio, S., Bourlard, H. (eds.) MLMI 2004. LNCS, vol. 3361, pp. 1–11. Springer, Heidelberg (2005)
14. Wellner, P., Flynn, M., Guillemot, M.: Browsing recorded meetings with Ferret. In: Bengio, S., Bourlard, H. (eds.) MLMI 2004. LNCS, vol. 3361, pp. 12–21. Springer, Heidelberg (2005)
15. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., Wellner, P.: The AMI meeting corpus: A pre-announcement. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, pp. 28–39. Springer, Heidelberg (2006)

16. Murray, G., Renals, S.: Term-weighting for summarization of multi-party spoken dialogues. In: Popescu-Belis, A., Renals, S., Bourlard, H. (eds.) MLMI 2007. LNCS, vol. 4892, pp. 155–166. Springer, Heidelberg (2008)
17. Kleinbauer, T., Becker, S., Becker, T.: Combining multiple information layers for the automatic generation of indicative meeting abstracts. In: Proc. of ENLG 2007, Dagstuhl, Germany (2007)