Ecological validity and the evaluation of speech summarization quality

McCallum, Anthony; Munteanu, Cosmin; Penn, Gerald; Zhu, Xiaodan

Publisher's version / Version de l'éditeur:

Proceeding of the Workshop on Evaluation Metrics and System Comparison for Automatic Summarization, pp. 28-35, 2012-07

Ecological Validity and the Evaluation of Speech Summarization Quality

Anthony McCallum
University of Toronto
40 St. George Street
Toronto, ON, Canada
mccallum@cs.toronto.edu

Cosmin Munteanu
National Research Council Canada
46 Dineen Drive
Fredericton, NB, Canada
cosmin.munteanu@nrc-cnrc.gc.ca

Gerald Penn
University of Toronto
40 St. George Street
Toronto, ON, Canada
gpenn@cs.toronto.edu

Xiaodan Zhu
National Research Council Canada
1200 Montreal Road
Ottawa, ON, Canada
xiaodan.zhu@nrc-cnrc.gc.ca

Abstract

There is little evidence of widespread adoption of speech summarization systems. This may be due in part to the fact that the natural language heuristics used to generate summaries are often optimized with respect to a class of evaluation measures that, while computationally and experimentally inexpensive, rely on subjectively selected gold standards against which automatically generated summaries are scored. This evaluation protocol does not take into account the usefulness of a summary in assisting the listener in achieving his or her goal.

In this paper we study how current measures and methods for evaluating summarization systems compare to human-centric evaluation criteria. For this, we have designed and conducted an ecologically valid evaluation that determines the value of a summary when embedded in a task, rather than how closely a summary resembles a gold standard. The results of our evaluation demonstrate that in the domain of lecture summarization, the well-known baseline of maximal marginal relevance (Carbonell and Goldstein, 1998) is statistically significantly worse than human-generated extractive summaries, and even worse than having no summary at all in a simple quiz-taking task. Priming seems to have no statistically significant effect on the usefulness of the human summaries. In addition, ROUGE scores and, in particular, the context-free annotations that are often supplied to ROUGE as references, may not always be reliable as inexpensive proxies for ecologically valid evaluations. In fact, under some conditions, relying exclusively on ROUGE may even lead to scoring human-generated summaries that are inconsistent in their usefulness relative to using no summaries very favourably.

1 Background and Motivation

Summarization maintains a representation of an entire spoken document, focusing on those utterances (sentence-like units) that are most important and therefore does not require the user to process everything that has been said. Our work focuses on extractive summarization where a selection of utterances is chosen from the original spoken document in order to make up a summary.

Current speech summarization research has made extensive use of intrinsic evaluation measures such as F-measure, Relative Utility, and ROUGE (Lin, 2004), which score summaries against subjectively selected gold standard summaries obtained using human annotators. These annotators are asked to arbitrarily select (in or out) or rank utterances, and in doing so commit to relative salience judgements with no attention to goal orientation and no requirement to synthesize the meanings of larger units of structure into a coherent message.

Given this subjectivity, current intrinsic evaluation measures are unable to properly judge which summaries are useful for real-world applications. For example, intrinsic evaluations have failed to show that summaries created by algorithms based on complex linguistic and acoustic features are better than baseline summaries created by simply choosing the positionally first utterances or longest utterances in a spoken document (Penn and Zhu, 2008). What is needed is an ecologically valid evaluation that determines how valuable a summary is when embedded in a task, rather than how closely a summary matches the subjective utterance level scores assigned by annotators.

Ecological validity is "the ability of experiments to tell us how real people operate in the real world" (Cohen, 1995). This is often obtained by using human judges, but it is important to realize that the mere use of human subjects provides no guarantee as to the ecological validity of their judgements. When utterances are merely ranked with numerical scores out of context, for example, the human judges who perform this task are not performing a task that they generally perform in their daily lives, nor does the task correspond to how they would create or use a good summary if they did have a need for one. In fact, there may not even be a guarantee that they *understand* the task --- the notions of "importance," "salience" and the like, when defining the criterion by which utterances are selected, are not easy to circumscribe. Judgements obtained in this fashion are no better than those of the generative linguists who leaned back in their armchairs in the 1980s to introspect on the grammaticality of natural language sentences. The field of computational linguistics could only advance when corpora became electronically available to investigate language that was written in an ecologically valid context.

Ours is not the first ecologically valid experiment to be run in the context of speech summarization, however. He et al. (1999; 2000) conducted a very thorough and illuminating study of speech summarization in the lecture domain that showed (1) speech summaries are indeed very useful to have around, if they are done properly, and (2) abstractive summaries do not seem to add any statistically significant advantage to the quality of a summary over what topline extractive summaries can provide. This is very good news; extractive summaries are worth creating. Our study extends this work by attempting to evaluate the relative quality of extractive summaries. We conjecture that it would be very difficult for this field to progress unless we have a means of accurately measuring extractive summarization quality. Even if the measure comes at great expense, it is important to do.

Another noteworthy paper is that of Liu and Liu (2010), who, in addition to collecting human summaries of six meetings, conducted a subjective assessment of the quality of those summaries with numerically scored questionnaires. These are known as *Likert scales*, and they form an important component of any human-subject study in the field of human-computer interaction. Liu and Liu (2010) cast considerable doubt on the value of ROUGE relative to these questionnaires. We will focus here on an objective, task-based measure that typically complements those subjective assessments.

2 Spontaneous Speech

Spontaneous speech is often not linguistically well-formed, and contains disfluencies, such as false starts, filled pauses, and repetitions. Additionally, spontaneous speech is more vulnerable to automatic speech recognition (ASR) errors, resulting in a higher word error rate (WER). As such, speech summarization has the most potential for domains consisting of spontaneous speech (e.g. lectures, meeting recordings). Unfortunately, these domains are not easy to evaluate compared to highly structured domains such as broadcast news. Furthermore, in broadcast news, nearly perfect studio acoustic conditions and professionally trained readers results in low ASR WER, making it an easy domain to summarize. The result is that most research has been conducted in this domain. However, a positional baseline performs very well in summarizing broadcast news (Christensen, 2004), meaning that simply taking the first N utterances provides a very challenging baseline, questioning the value of summarizing this domain. In addition, the widespread availability of written sources on the same topics means that there is not a strong use case for speech summarization over simply summarizing the equivalent textual articles on which the news broadcasts were based. This makes it even more difficult to preserve ecological validity.

University lectures present a much more relevant domain, with less than ideal acoustic conditions but structured presentations in which deviation

from written sources (e.g., textbooks) is commonplace. Here, a positional baseline performs very poorly. The lecture domain also lends itself well to a task-based evaluation measure; namely university level quizzes or exams. This constitutes a real-world problem in a domain that is also representative of other spontaneous speech domains that can benefit from speech summarization.

3 Ecologically Valid Evaluation

As pointed out by Penn and Zhu (2008), current speech summarizers have been optimized to perform an utterance selection task that may not necessarily reflect how a summarizer is able to capture the goal orientation or purpose of the speech data. In our study, we follow methodologies established in the field of Human-Computer Interaction (HCI) for evaluating an algorithm or system – that is, determining the benefits a system brings to its users, namely usefulness, usability, or utility, in allowing a user to reach a specific goal. Increasingly, such user-centric evaluations are carried out within various natural language processing applications (Munteanu et al., 2006). The prevailing trend in HCI is for conducting extrinsic summary evaluations (He et al., 2000; Murray et al., 2008; Tucker et al., 2010), where the value of a summary is determined by how well the summary can be used to perform a specific task rather than comparing the content of a summary to an artificially created gold standard. We have conducted an ecologically valid evaluation of speech summarization that has evaluated summaries under real-world conditions, in a task-based manner.

The university lecture domain is an example of a domain where summaries are an especially suitable tool for navigation. Simply performing a search will not result in the type of understanding required of students in their lectures. Lectures have topics, and there is a clear communicative goal. For these reasons, we have chosen this domain for our evaluation. By using actual university lectures as well as university students representative of the users who would make use of a speech summarization system in this domain, all results obtained are ecologically valid.

3.1 Experimental Overview

We conducted a within-subject experiment where participants were provided with first year sociology university lectures on a lecture browser system installed on a desktop computer. For each lecture, the browser made accessible the audio, manual transcripts, and an optional summary. Evaluation of a summary was based on how well the user of the summary was able to complete a quiz based on the content of the original lecture material.

It is important to note that not all extrinsic evaluation is ecologically valid. To ensure ecological validity in this study, great care was taken to ensure that human subjects were placed under conditions that result in behavior that would be expected in actual real-world tasks.

3.2 Evaluation

Each quiz consisted of 12 questions, and were designed to be representative of what students were expected to learn in the class, incorporating factual questions only, to ensure that variation in participant intelligence had a minimal impact on results. In addition, questions involved information that was distributed equally throughout the lecture, but at the same time not linearly in the transcript or audio slider, which would have allowed participants to predict where the next answer might be located. Finally, questions were designed to avoid content that was thought to be common knowledge in order to minimize the chance of participants having previous knowledge of the answers.

All questions were short answer or fill-in-the-blank. Each quiz consisted of an equal number of four distinct types of questions, designed so that performing a simple search would not be effective, though no search functionality was provided. Question types do not appear in any particular order on the quiz and were not grouped together.

Type 1: These questions can be answered simply by looking at the slides. As such, these questions could be answered correctly with or without a summary as slides were available in all conditions.

Type 2: Slides provide an indication of where the content required to answer these questions are located. Access to the corresponding utterances is still required to find the answer to the questions.

Type 3: Answers to these questions can only be found in the transcript and audio. The slides provide no hint as to where the relevant content is located.

Type 4: These questions are more complicated and require a certain level of topic comprehension. These questions often require connecting concepts from various portions of the lecture. These questions are more difficult and were included to minimize the chance that participants would already know the answer to questions without watching the lecture.

A teaching assistant for the sociology class from which our lectures were obtained generated the quizzes used in the evaluation. This teaching assistant had significant experience in the course, but was not involved in the design of this study and did not have any knowledge relating to our hypotheses or the topic of extractive summarization. These quizzes provided an ecologically valid quantitative measure of whether a given summary was useful. Having this evaluation metric in place, automated summaries were compared to manual summaries created by each participant in a previous session.

3.3 Participants

Subjects were recruited from a large university campus, and were limited to undergraduate students who had at least two terms of university studies, to ensure familiarity with the format of university-level lectures and quizzes. Students who had taken the first year sociology course we drew lectures from were not permitted to participate. The study was conducted with 48 participants over the course of approximately one academic semester.

3.4 Method

Each evaluation session began by having a participant perform a short warm-up with a portion of lecture content, allowing the participant to become familiar with the lecture browser interface. Following this, the participant completed four quizzes, one for each of four lecture-condition combinations. There were a total of four lectures and four conditions. Twelve minutes were given for each quiz. During this time, the participant was able to browse the audio, slides, and summary. Each lecture was about forty minutes in length, establishing a time constraint. Lectures and conditions were rotated using a Latin square for counter balancing. All participants completed each of the four conditions.

One week prior to his or her evaluation session, each participant was brought in and asked to listen to and summarize the lectures beforehand. This resulted in the evaluation simulating a scenario where someone has heard a lecture at least one week in the past and may or may not remember the content during an exam or quiz. This is similar to conditions most university students experience.

3.5 Conditions

The lecture audio recordings were manually transcribed and segmented into utterances, determined by 200 millisecond pauses, resulting in segments that correspond to natural sentences or phrases. The task of summarization consisted of choosing a set of utterances for inclusion in a summary (extractive summarization), where the total summary length was bounded by 17-23% of the words in the lecture; a percentage typical to most summarization scoring tasks. All participants were asked to make use of the browser interface for four lectures, one for each of the following conditions: *no summary*, *generic manual summary*, *primed manual summary*, and *automatic summary*.

No summary: This condition served as a baseline where no summary was provided, but participants had access to the audio and transcript. While all lecture material was provided, the twelve-minute time constraint made it impossible to listen to the lecture in its entirety.

Generic manual summary: I this condition, each participant was provided with a manually generated summary. Each summary was created by the participant him or herself in a previous session. Only audio and text from the in-summary utterances were available for use. This condition demonstrates how a manually created summary is able to aid in the task of taking a quiz on the subject matter.

Primed manual summary: Similar to above, in this condition, a summary was created manually by selecting a set of utterances from the lecture transcript. For primed summaries, full access to a priming quiz, containing all of the questions in the evaluation quiz as well as several additional questions, was available at the time of summary cre-

ation. This determines the value of creating summaries with a particular task in mind, as opposed to simply choosing utterances that are felt to be most important or salient.

Automatic summary: The procedure for this condition was identical to the *generic manual summary* condition from the point of view of the participant. However, during the evaluation phase, an automatically generated summary was provided instead of the summary that the participant created him or herself. The algorithm used to generate this summary was an implementation of MMR (Carbonell and Goldstein, 1998). Cosine similarity with tf-idf term weighting was used to calculate similarity. Although the redundancy component of MMR makes it especially suitable for multi-document summarization, there is no negative effect if redundancy is not an issue. It is worth noting that our lectures are longer than material typically summarized, and lectures in general are more likely to contain redundant material than a domain such as broadcast news. There was only one MMR summary generated for each lecture, meaning that multiple participants made use of identical summaries. The automatic summary was created by adding the highest scoring utterances one at a time until the sum of the length of all of the selected utterances reached 20% of the number of words in the original lecture. MMR was chosen as it is commonly used in summarization. MMR is a competitive baseline, even among state-of-art summarization algorithms, which tend to correlate well with it.

What this protocol does not do is pit several strategies for automatic summary generation against each other. That study, where more advanced summarization algorithms will also be examined, is forthcoming. The present experiments have the collateral benefit of serving as a means for collecting ecologically valid human references for that study.

3.6 Results

Quizzes were scored by a teaching assistant for the sociology course from which the lectures were taken. Quizzes were marked as they would be in the actual course and each question was graded with equal weight out of two marks. The scores were then converted to a percentage. The resulting scores (Table 1) are 49.3+-17.3% for the *no summary* condition, 48.0+-16.2% for the *generic manual summary* condition, 49.1+-15.2% for the *primed summary* condition, and 41.0+-16.9% for *MMR*. These scores are lower than averages expected in a typical university course. This can be partially attributed to the existence of a time constraint.

Condition	Average Quiz Score
no summary	49.3+-17.3%
generic manual summary	48.0+-16.2%
primed manual summary	49.1+-15.2%
automatic summary (MMR)	41.0+-16.9%

Table 1. Average Quiz Scores

Execution of the Shapiro-Wilk Test confirmed the scores are normally distributed and Mauchly's Test of Sphericity indicates that the sphericity assumption holds. Skewness and Kurtosis tests were also employed to confirm normality. A repeated measures ANOVA determined that scores varied significantly between conditions ($F(3,141)=5.947$, $P=0.001$). Post-hoc tests using the Bonferroni correction indicate that the *no summary*, *generic manual summary*, and *primed manual summary* conditions all resulted in higher scores than the *automatic (MMR) summary condition*. The difference is significant at $P=0.007$, $P=0.014$ and $P=0.012$ respectively. Although normality was assured, the Friedman Test further confirms a significant difference between conditions ($\chi^2(3)=11.684$, $P=0.009$).

4 F-measure

F-measure is an evaluation metric that balances precision and recall which has been used to evaluate summarization. Utterance level F-measure scores were calculated using the same summaries used in our human evaluation. In addition, three annotators were asked to create conventional gold standard summaries using binary selection. Annotators were not primed in any sense, did not watch the lecture videos, and had no sense of the higher level purpose of their annotations. We refer to the resulting summaries as context-free as they were not created under ecologically valid conditions. F-measure was also calculated with reference to these.

The F-measure results (Table 2) point out a few interesting phenomena. Firstly, when evaluating a

given peer summary type with the same model type, the *generic-generic* scores are higher than both the *primed-primed* and *context-free-context-free* summaries. This means that generic summaries tend to share more utterances with each other, than primed summaries do, which are more varied. This seems unintuitive at first, but could potentially be explained by the possibility that different participants focused on different aspects of the priming quiz, due to either perceived importance, or lack of time (or summary space) to address all of the priming questions.

Peer Type	Model Type	Average F-measure
generic	*generic*	0.388
primed	*generic*	0.365
MMR	*generic*	0.214
generic	*primed*	0.365
primed	*primed*	0.374
MMR	*primed*	0.209
generic	*context-free*	0.371
primed	*context-free*	0.351
MMR	*context-free*	0.243
context-free	*context-free*	0.374

Table 2. Average F-measure

We also observe that generic summaries are more similar to conventionally annotated (context-free) summaries than either primed or MMR are. This makes sense and also confirms that even though primed summaries do not significantly outperform generic summaries in the quiz taking task, they are inherently distinguishable from each other.

Furthermore, when evaluating MMR using F-measure, we see that MMR summaries are most similar to the context-free summaries, whose utterance selections can be considered somewhat arbitrary. Our quiz results confirm MMR is significantly worse than generic and primed summaries. This casts doubt on the practice of using similarly annotated summaries as gold standards for summarization evaluation using ROUGE.

5 ROUGE Evaluation

More common than F-measure, ROUGE (Lin, 2004) is often used to evaluate summarization. Although Lin (2004) claimed to have demonstrated that ROUGE correlates well with human summaries, both Murray et al. (2005), and Liu and Liu (2010) have cast doubt upon this. It is important to acknowledge, however, that ROUGE is actually a family of measures, distinguished not only by the manner in which overlap is measured (1-grams, longest common subsequences, etc.), but by the provenience of the summaries that are provided to it as references. If these are not ecologically valid, there is no sense in holding ROUGE accountable for an erratic result.

To examine how ROUGE fairs under ecologically valid conditions, we calculated ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4 on our data using the standard options outlined in previous DUC evaluations. ROUGE scores were calculated for each of the *generic manual summary*, *primed manual summary*, and *automatic summary* conditions. Each summary in a given condition was evaluated once against the *generic manual* summaries and once using the *primed manual* summaries. Similar to Liu and Liu (2010), ROUGE evaluation was conducted using leave-one-out on the model summary type and averaging the results.

In addition to calculating ROUGE on the summaries from our ecologically valid evaluation, we also followed more conventional ROUGE evaluation and used the same context-free annotator summaries as were used in our F-measure calculations above. Using these context-free summaries, the original generic manual, primed manual, and automatic summaries were evaluated using ROUGE. The result of these evaluations are presented in Table 3.

Looking at the ROUGE scores, we can see that when evaluated by each type of model summary, MMR performs worse than either generic or primed manual summaries. This is consistent with our quiz results, and perhaps shows that ROUGE may be able to distinguish human summaries from MMR. Looking at the *generic-generic*, *primed-primed*, and *context-free-context-free* scores, we can get a sense of how much agreement there was between summaries. It is not surprising that context-free annotator summaries showed the least agreement, as these summaries were generated with no higher purpose in mind. This suggests that using annotators to generate gold standards in such a manner is not ideal. In addition, real world applications for summarization would conceivably

rarely consist of a situation where a summary was created for no apparent reason. More interesting is the observation that, when measured by ROUGE, primed summaries have less in common with each other than generic summaries do. The difference, however, is less pronounced when measured by ROUGE than by F-measure. This is likely due to the fact that ROUGE can account for semantically similar utterances.

Peer type	Model type	R-1	R-2	R-L	R-SU4
generic	*generic*	0.75461	0.48439	0.75151	0.51547
primed	*generic*	0.74408	0.46390	0.74097	0.49806
MMR	*generic*	0.71659	0.40176	0.71226	0.44838
generic	*primed*	0.74457	0.46432	0.74091	0.49844
primed	*primed*	0.74693	0.46977	0.74344	0.50254
MMR	*primed*	0.70773	0.38874	0.70298	0.43802
generic	*context-free*	0.72735	0.46421	0.72432	0.49573
primed	*context-free*	0.71793	0.44325	0.71472	0.47805
MMR	*context-free*	0.69233	0.37600	0.68813	0.42413
context-free	*context-free*	0.70707	0.44897	0.70365	0.48019

Table 3. Average ROUGE Scores

5.1 Correlation with Quiz Scores

In order to assess the ability of ROUGE to predict quiz scores, we measured the correlation between ROUGE scores and quiz scores on a per participant basis. Similar to Murray et al. (2005), and Liu and Liu (2010), we used Spearman's rank coefficient (rho) to measure the correlation between ROUGE and our human evaluation. Correlation was measured both by calculating Spearman's rho on all data points ("all" in Table 4) and by performing the calculation separately for each lecture and averaging the results ("avg"). Significant rho values (p-value less than 0.05) are shown in bold.

Note that there are not many bolded values, indicating that there are few (anti-)correlations between quiz scores and ROUGE. The rho values reported by Liu and Liu (2010) correspond to the "all" row of our generic-context-free scores (Liu and Liu (2010) did not report ROUGE-L), and we obtained roughly the same scores as they did. In contrast to this, our "all" generic-generic correlations are very low. It is possible that the lectures condition the parameters of the correlation to such an extent that fitting all of the quiz-ROUGE pairs to the same correlation across lectures is unreasonable. It may therefore be more useful to look at rho values computed by lecture. For these values, our R-SU4 scores are not as high relative to R-1 and R-2 as those reported by Liu and Liu (2010). It is also worth noting that the use of context-free binary selections as a reference results in increased correlation for generic summaries, but substantially decreases correlation for primed summaries.

With the exception that generic references prefer generic summaries and primed references prefer primed summaries, all other values indicate that both generic and primed summaries are better than MMR. However, instead of ranking summary types, what is important here is the ecologically valid quiz scores. Our data provides no evidence that ROUGE scores accurately predict quiz scores.

6 Conclusions

We have presented an investigation into how current measures and methodologies for evaluating summarization systems compare to human-centric evaluation criteria. An ecologically-valid evaluation was conducted that determines the value of a summary when embedded in a task, rather than how closely a summary resembles a gold standard. The resulting quiz scores indicate that manual summaries are significantly better than MMR. ROUGE scores were calculated using the summaries created in the study. In addition, more conventional context-free annotator summaries were also used in ROUGE evaluation. Spearman's rho indicated no correlation between ROUGE scores and our ecologically valid quiz scores. The results offer evidence that ROUGE scores and particularly context-free annotator-generated summaries as gold standards may not always be reliably used in place of an ecologically valid evaluation.

Peer type	Model type		R-1	R-2	R-L	R-SU4
generic	*generic*	all	0.017	0.066	0.005	0.058
		lec1	0.236	0.208	0.229	0.208
		lec2	0.276	0.28	0.251	0.092
		lec3	0.307	**0.636**	0.269	0.428
		lec4	0.193	-0.011	0.175	0.018
		avg	0.253	0.278	0.231	0.187
primed	*generic*	all	-0.097	-0.209	-0.090	-0.192
		lec1	-0.239	-0.458	-0.194	-0.458
		lec2	-0.306	-0.281	-0.306	-0.316
		lec3	0.191	0.142	0.116	0.255
		lec4	**-0.734**	**-0.78**	**-0.769**	**-0.78**
		avg	-0.272	-0.344	-0.288	-0.325
generic	*primed*	all	0.009	0.158	-0.004	0.133
		lec1	0.367	0.247	0.367	0.162
		lec2	**0.648**	0.425	**0.634**	0.304
		lec3	0.078	0.417	0.028	0.382
		lec4	0.129	0.079	0.115	0.025
		avg	0.306	0.292	0.286	0.218
primed	*primed*	all	0.161	0.042	0.161	0.045
		lec1	0.042	-0.081	0.042	-0.194
		lec2	0.238	0.284	0.259	0.284
		lec3	0.205	0.12	0.205	0.12
		lec4	0.226	0.423	0.314	0.423
		avg	0.178	0.187	0.205	0.158
generic	*con-text-free*	all	0.282	**0.306**	0.265	**0.347**
		lec1	-0.067	0.296	-0.004	0.325
		lec2	0.414	0.414	0.438	0.319
		lec3	0.41	0.555	0.41	0.555
		lec4	0.136	0.007	0.136	0.054
		avg	0.223	0.318	0.245	0.313
primed	*con-text-free*	all	-0.146	-0.282	-0.151	**-0.305**
		lec1	0.151	-0.275	0.151	-0.299
		lec2	-0.366	**-0.611**	-0.366	**-0.636**
		lec3	0.273	0.212	0.273	0.202
		lec4	**-0.815**	-0.677	**-0.825**	**-0.755**
		avg	-0.189	-0.338	-0.192	-0.372

Table 4. Correlation (Spearman's rho) between Quiz Scores and ROUGE

7 References

J. Carbonell and J. Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 335-336, ACM.

P. R. Cohen. 1995. *Empirical methods for artificial intelligence*. Volume 55. MIT press Cambridge, Massachusetts.

H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals. 2004. From text summarisation to style-specific summarisation for broadcast news. *Advances in Information Retrieval*, 223-237.

L. He, E. Sanocki, A. Gupta, and J. Grudin. 1999. Auto-summarization of audio-video presentations. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, 489-498. ACM.

L. He, E. Sanocki, A. Gupta, and J. Grudin. 2000. Comparing presentation summaries: slides vs. reading vs. listening. In *Proc. of the SIGCHI*, 177-184, ACM.

C. Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Proc. of ACL, Text Summarization Branches Out Workshop*, 74–81.

F. Liu and Y. Liu. 2010. Exploring correlation between rouge and human evaluation on meeting summaries. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(1):187-196.

C. Munteanu, R. Baecker, G. Penn, E. Toms, and D. James. 2006. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 493-502, ACM.

G. Murray, S. Renals, J. Carletta, and J. Moore. 2005. Evaluating automatic summaries of meeting recordings. In *Proc. of the ACL 2005 MTSE Workshop*, Ann Arbor, MI, USA, 33-40.

G. Murray, T. Kleinbauer, P. Poller, S. Renals, J. Kilgour, and T. Becker. 2008. Extrinsic summarization evaluation: A decision audit task. *Machine Learning for Multimodal Interaction*, 349-361.

G. Penn and X. Zhu. 2008. A critical reassessment of evaluation baselines for speech summarization. *Proc. of ACL-HLT*.

S. Tucker, O. Bergman, A. Ramamoorthy, and S. Whittaker. 2010. Catchup: a useful application of time-travel in meetings. In *Proc. of CSCW*, 99-102, ACM.

S. Tucker and S. Whittaker. 2006. Time is of the essence: an evaluation of temporal compression algorithms. In *Proc. of the SIGCHI*, 329-338, ACM.