An Ecologically Valid Evaluation of Speech Summarization

Anthony McCallum
University of Toronto
40 St. George Street
Toronto, ON M5S 2E4 CANADA
mccallum@cs.toronto.edu

Cosmin Munteanu
National Research Council Canada
46 Dineen Dr.
Fredericton, NB E3B 9W4 CANADA
cosmin.munteanu@nrc-cnrc.gc.ca

Gerald Penn
University of Toronto
40 St. George Street
Toronto, ON M5S 2E4 CANADA
gpenn@cs.toronto.edu

Xiaodan Zhu
National Research Council Canada
1200 Montreal Road
Ottawa, ON, K1A 0R6 CANADA
xiaodan.zhu@nrc-cnrc.gc.ca

CHI 2012, May 5–10, 2012, Austin, TX, USA.
ACM xxx-x-xxxx-xxxx-x/xx/xx.

Abstract

The past decade has witnessed an explosion in the size and availability of online audio-visual repositories, such as entertainment, news, or lectures. Summarization systems have the potential to provide significant assistance with navigating such repositories. Unfortunately, automatically-generated summaries often fall short of delivering the information needed by users. This is due, in no small part, to the fact that the natural language heuristics used to generate summaries are often optimized with respect to currently-used evaluation metrics. Such metrics simply score automatically-generated summaries against subjectively-classified gold standards without taking into account the usefulness of a summary in assisting a user achieve a certain goal or even overall summary coherence. We have previously shown that an immediate consequence of this problem is that even the most linguistically-complex summarization systems perform no better than basic heuristics, such as picking the longest sentences from a general-topic, spontaneous dialog, or the first few sentences from a news recording. Our hypothesis is that complex systems are in fact better, if measured properly. What is thus needed instead are evaluation metrics (and consequently, automatic summarizers) that incorporate features such as user preferences and task-orientation. For this, we propose an ecologically valid evaluation metric that determines the value of a summary when embedded in a task, rather than how closely a summary matches a gold standard.

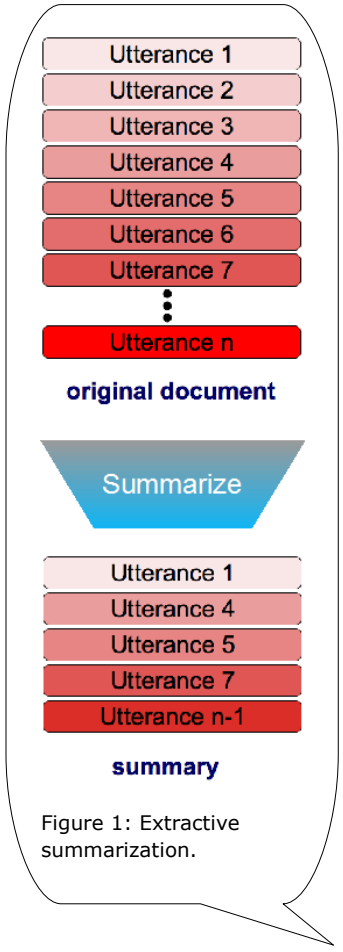


Figure 1: Extractive summarization.

Author Keywords
Speech summarization; user study; extrinsic evaluation

ACM Classification Keywords
L.2.7 [Natural language processing]: Speech recognition and synthesis;

General Terms
Experimentation, human factors, performance.

Background and motivation

As one of the most natural means of communication between humans, it is not surprising that speech continues to be used effectively in a wide variety of applications. Widespread availability of recording equipment, affordable digital storage, and the pervasiveness of high bandwidth Internet connectivity has resulted in an ever-increasing amount of audio-visual material at our fingertips. Companies archive past conference calls and meetings, and universities often make recordings of lectures available online. An astounding 48 hours of video is uploaded to YouTube each minute [8]. Not surprisingly, the availability of methods to navigate speech has become essential, as it is impossible and unnecessary to view all relevant speech in its entirety in order to extract desired information.

The solution we choose to focus on is speech summarization [7]. 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. Current research focuses on extractive summarization where a selection of utterances is chosen from the original spoken document in order to make up a summary. Extractive summaries can be generated using any raw audio source, without the need for human transcription, by relying on automatic speech recognition (ASR) or simply using acoustic features. Extractive summaries can be displayed visually as text, or also audially. The latter may be most effective in situations with a relatively high ASR word error rate (WER), allowing users to listen to the original audio where reading an error-laden transcript may be unfavourable.

Current speech summarization research makes use of intrinsic evaluation metrics such as F-measure, Relative Utility and ROUGE [3], which score summaries against subjectively classified gold standard summaries obtained using human annotators. Annotators are asked to rank utterances on a numerical scale, 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. The utterances in automatically generated summaries are then evaluated by either summing the annotator-assigned scores or by counting the number of annotator-selected utterances that appear in the automatic summary.

Given this subjectivity, current intrinsic evaluation metrics are unable to properly judge which summaries are useful for real world applications. For example, using these intrinsic measures has 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 first utterances occurring in a spoken document or the longest utterances [5]. What is needed is an ecologically valid evaluation metric 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. It is possible that our study will demonstrate that current state-of-the-art automatic summarizers, particularly those which rely on linguistically rich features, will be far more effective, when evaluated using an extrinsic evaluation task, than simple baselines. In the event of such an outcome, our now ecologically validated manual summaries can be used as a sort of gold standard for future optimizations of automatic speech summarizers.

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 ASR errors, resulting in higher 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 [1], 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.

University lectures present a much more relevant domain, with less than ideal acoustic conditions and a structured presentation in which deviation from written sources (e.g., textbooks) is commonplace, and yet a positional baseline performs very poorly. The lecture domain also lends itself well to a task-based evaluation metric; 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.

Task-orientated evaluation

As pointed out by [5], 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 the prevailing trend in HCI towards extrinsic summary evaluation, 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 [4,6].

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. 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, any results obtained will be ecologically valid.

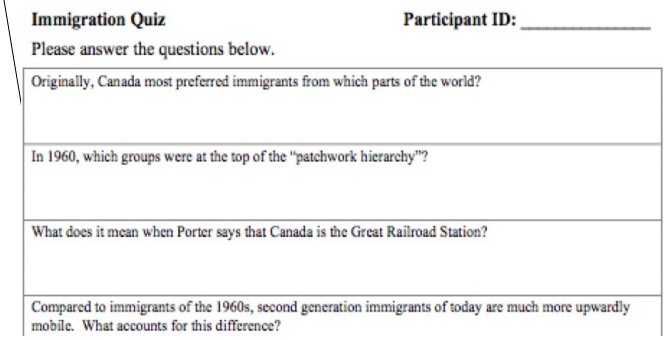
The 12-question quizzes have been designed to be representative of what students are expected to learn in the class, incorporating factual questions only to ensure that variation in participant intelligence has a minimal impact on results. In addition, questions involve information that is distributed equally throughout the lecture, but at the same time not linearly in the transcript or audio slider, which would allow participants to predict where the next answer might be located. Finally, all questions are non-trivial to minimize the chance of the participant having previous knowledge of the answer. We also included a mechanism for assessment of previous knowledge that we use to normalize the final scores.

Experimental design

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

Evaluation

A teaching assistant for the sociology class from which our lectures were obtained generated the quizzes used in the evaluation (Figure 2). These quizzes provide an ecologically valid quantitative measure of whether a given summary is useful. Having this evaluation metric in place, automated summaries are compared to manual summaries created by human summarizers familiar with the quiz content as well as summaries where the summarizer has not previously seen the evaluation quiz. This allows us to demonstrate the value of primed summarization, as well as determine which utterances an ideal summary would consist of.



Immigration Quiz **Participant ID: _____**

Please answer the questions below.

Originally, Canada most preferred immigrants from which parts of the world?
In 1960, which groups were at the top of the "patchwork hierarchy"?
What does it mean when Porter says that Canada is the Great Railroad Station?
Compared to immigrants of the 1960s, second generation immigrants of today are much more upwardly mobile. What accounts for this difference?

Figure 2: Example of an evaluation quiz.

Participants are also asked to complete questionnaires used to elicit qualitative data about their experience with using summaries in the quiz-taking task.

Participants

Subjects are recruited from a large university campus. Participants are limited to undergraduate students who have had at least two terms of university studies, to ensure familiarity with the format of university-level lectures and quizzes. Students who have taken the first year sociology course that we have drawn lectures from are not permitted to participate. The study is carried out with 96 participants as well as 4 participants used for normalization of quiz scores across lectures.

Method

A given session begins 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 (Figure 3). Following this, the participant completes four quizzes, one for each of four lecture-condition combinations. There are a total of four lectures and four conditions. Twelve minutes are given for each quiz. During this time, the participant is able to browse the audio, slides, and summary. Each lecture is about forty minutes in length, establishing a strong time constraint that could make using a well-prepared summary beneficial to completing the task at hand.

Lectures and conditions are rotated using a Latin square for counter balancing. All participants complete each of the four conditions. However, half of the participants are exposed to the lecture material for the first time, preventing any possible recall effects, while the other half have already heard and summarized the lectures during a previous session. We are, in effect,

The interface provides a table of contents (slide thumbnails, on left), the current slide (top), an interactive timeline (middle), transcripts of either the full lecture or of the summary, depending on the experimental condition (bottom centre), and optionally, the summarization creation tool (allowing drag-and-drop manipulation of sentences – on the right side). Previous research ([2]) suggests that embedding summaries in lecture interfaces improves users' performance and experience in information-seeking tasks.

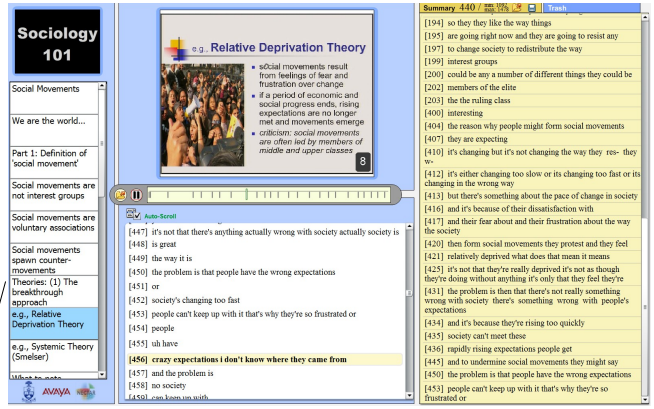


Figure 3: The lecture browsing interface.

conducting two studies with the same experimental protocol, but separate subjects, where subjects in the first study have been previously exposed to all lectures. We are thus simulating both a scenario in which someone wants to extract information from a lecture that he or she has not previously heard, as well as a scenario where someone has heard a lecture at least one week in the past and may or may not remember the content. For those conditions that rely on a manual summary, a separate human summarizer is used to create this summary for the participants who have not previously seen the lecture, while participants who have already seen the lectures use self-created summaries.

Conditions

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

The *no summary* condition serves as a baseline where no summary is provided, but participants have access to the audio and transcript. While all lecture material is provided, the twelve-minute time constraint makes it impossible to listen to the lecture in its entirety.

In the *automatic summary* condition, the summary is generated using the automatic summarizer described in [5], correlating to, and performing at least as well as the commonly used Maximal Marginal Relevance (MMR) algorithm. This summarizer makes use of a wide variety of speech features and is representative of the state-of-the-art. The summarizer takes either ASR or manual transcripts as well as an audio file as input that it uses to process disfluencies and extract various features important to identifying sentences. False starts and repetitions, which occur commonly in spontaneous speech, are detected and removed. A binary logistic regression classifier is used to train an utterance selection module that can make use of various lexical (MMR score, utterance length, etc.), structural (utterance position, etc.), and acoustic (pitch, energy, speaking rate, etc.) features, among others. Given an ecologically valid experimental setup and evaluation metric, results obtained for this condition may confirm that current state-of-the-art summarizers do indeed perform better than a simple length baseline on lecture

data, a baseline that under previous intrinsic evaluation conditions has been very competitive [5].

In the *generic manual summary* condition, each participant is provided with a manually generated summary. Each summary is created either by the participant herself in a previous session or by a separate human summarizer who has listened to the lecture in its entirety. This condition demonstrates how a manually created summary is able to aid in the task of taking a quiz on the subject matter. Along with the next condition, this summary provides a comparison for which to evaluate the performance of current state-of-the-art automatic summarizers against.

Similar to above, in the *primed manual summary* condition, a summary is created manually by selecting a set of utterances from the lecture transcript. For primed summaries, full access to the evaluation quiz is available at the time of summary creation. 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. If such summaries result in participants performing the task well, then we will be able use these summaries to gain a better understanding of what an ideal summary should contain.

Progress to date and Conclusions

We have proposed an ecologically valid evaluation of speech summarization using the university lecture domain. We will evaluate the value of primed summaries as well as use a task-based metric for determining the value of a summary. The resulting verified summaries will set a new high-water mark for evaluation within spoken language processing research. The study has been conducted using the first set of 48 participants, where lectures were viewed at least one week prior to the evaluation and manual summaries were created by the participants themselves. Outstanding work involves conducting the experiment with the remaining 48 participants, where lectures are unseen and summaries are not self-created, as well as completing the qualitative and quantitative data collection and analysis.

References

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

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

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

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

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

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

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

[8] YouTube. http://www.youtube.com/t/press_statistics