UNKNOWN

Kathleen McKeown and Julia Hirschberg

Columbia University
New York NY 10027
{kathy, julia}@cs.columbia.edu

ABSTRACT

1. INTRODUCTION

How is summarization done for text and why does this fall short for speech summarization

Problems in general with speech summarization

Additional information that can be useful for speech summarization not available for text.

Different kinds of speech and this influences what kind of summarization can be done.

2. APPROACHES TO SUMMARIZATION

Current summarization systems can be categorized by the type of input that they handle, by the approach used and by the type of summary that they generate. Single document summarization systems generate a summary of one document as input while multi-document summarizations generates a summary of a set of documents on the same topic or event. Statistical extraction systems generate a summary by lifting sentences from the input article and stringing them together, while abstractive systems attempt to synthesize new sentences. Human summarizers have long distinguished between informative summaries, those that convey the content of the article and can be read in place of the document, and indicative summaries, those that describe characteristics of the document (e.g., its topic, length, style) and can be used to determine if the document is of interest.

To allow summarization in arbitrary domains, most current systems use sentence extraction, identifying and extracting key sentences from an input article using a variety of different criteria. The key sentences are then strung together to form the summary. These approaches have all been developed to produce a summary of a single input document. Early approaches used statistical metrics (e.g., word frequencies and key phrases) to identify important sentences [**?**, **?**, **?**]. More recent approaches [**?**] use a corpus of articles with summaries for training to identify the features of sentences that are typically included in abstracts. Other recent approaches use lexical chains [**?**], sentence position [**?**], discourse structure [**?**, **?**], and user features from the query [**?**] to score sentences and label them as key. Problems for this approach center around accidentally including pronouns which have no previous reference in the extracted text (a problem addressed by [**?**]) or, in the case of extracting several sentences, of including incoherent text when the extracted sentences are not consecutive in the original text and do not naturally follow one another.

Extractive systems tend to produce summaries with very long sentences. That is because, in general, the longer sentences score higher on metrics that rate them for importance. Abstrative approaches to single document summarization address this problem by editing the sentences selected by extractive methods. The majority of this work focuses on compression. The aim is to reduce a sentence by eliminating constituents which are not crucial for its understanding nor salient enough to include in the summary. These approaches are based on the observation that the “importance” of a sentence constituent can often be determined based on shallow features, such as its syntactic role and the words it contains. For example, in many cases a relative clause that is peripheral to the central point of the document can be removed from a sentence without significantly distorting its meaning. While earlier approaches for text compression were based on symbolic reduction rules [**?**, **?**], more recent approaches use an aligned corpus of documents and their human written summaries to determine which constituents can be reduced [**?**, **?**, **?**]. Alignment is made between the summary sentences, which have been manually compressed, and the original sentences from which they were drawn.

Summarization across multiple documents has also often been addressed through sentence extraction. Many approaches generate a summary that focuses on similarities found across all articles; they use clustering to find common themes within the articles [**?**, **?**, **?**] producing sets of sentences where each set ,or *theme*, contains sentences saying roughly the same thing. Extractive approaches will extract one sentence from each set to form the summary. Other multi-document extractive approaches use information about the centroid of the documents [**?**] or lexical and structural information indicating importance [**?**] to find and extract key sentences. Mani and Bloedorn [**?**] use spreading activation and graph matching to compute similarities and differences between the salient topics of two articles. Output is presented as a set of paragraphs which contain similar and distinguishing words, emphasized in different fonts. The problem is a redundant summary since no synthesis of results through generation is attempted.

Only a few researchers have developed abstractive approaches for multi-document summarization. An approach based on information fusion [**?**, **?**] starts from the identification of themes as described above, but instead of extracting a representative sentence from the theme, uses alignment to find phrases that occur in multiple sentences within the theme. These phrases are extracted and a statistical language generation technique is used to fuse the phrases forming a novel sentence for the summary. Earlier work on multi-document summarization [**?**, **?**] used a symbolic approach, pairing information extraction with language generation. This type of approach produces more of a briefing than a summary. The system always looks for certain types of information (e.g., in a terrorist article, the event, the victims, the perpetrators, the location and the date) and generates a summary about this information regardless of the focus of the article. However, because it generates a summary from structured documents, it can highlight differences as well as similarities. The result is a domain dependent system for summarization of multiple news articles on the same event, highlighting how perspective of the event has changed over time.

3. SUMMARIZATION OF SPOKEN LANGUAGE

Speech summarization is a much harder task than text summarization. It may be difficult to identify utterance boundaries, utterances may be fragmentary and may contain disfluencies, and speech recognition may introduce additional errors. These characteristics mean that the extractive approaches used for text summarization will not necessarily work for speech summarization. We still need to be able to identify utterances that convey important content, but we must develop approaches that can substantially alter the extracted material in order to produce a good summary. Thus, it seems that speech summarization systems require an abstractive approach over a purely extractive one. Given these difficulties, summarization of spoken sources has, to date, included single document summarization only.

Speech summarization also has opportunities that do not exist for text summarizaiton. Information from the speech signal, such as prosody, can help a system to identify important content. Information about the speakers can also help determine importance; who is speaking, where the turn falls in relation to other speakers, and how the dialog is structured are important clues.

In the remainder of this paper, we describe ongoing research at Columbia towards summarization of two different types of speech sources.

3.1. Summarization of Broadcast News

While speech summarization techniques have been applied to genres such as recorded lectures, meetings, and voicemail, to date most speech summarization applications have focussed on Broadcast News [4, 2, 3, 5, 6]. Such data closely resembles the newswire data that much work in text summarization has concentrated on. Too, there is a large amount of training data available for study, and automatic speech recognition systems to provide transcriptions of reasonable accuracy. However, most of this research has assumed that such transcripts will be available and of high quality, on which techniques similar to text summarization techniques can then be employed. For example, [2] has used statistical methods to identify words to include in a summary, based upon linguistic features of the transcribed text, while [4] have used lexical extraction methods to hypothesize headlines for news programs. However, such methods are still limited by the quality of the speech transcription itself, especially in proper names, and other spontaneous speech phenomena, which make the approach of first transcribing into text and then using text-based summarization methods less than successful. To address this, [5] integrate the recognition process with a compression approach to summarization, pruning disfluencies during recognition, scoring the result based on acoustic confidence information as well as lexical likelihoods, and compressing the output to include only 'important' and well-recognized words.

In our work at Columbia on the summarization of Broadcast News [6], we pursued a **two-level** approach to the problem of summarizing errorful spoken material: First, we identify domain-specific aspects of newscasts to provide an **outline** of the newscast, which users can navigate in a GUI interface, following links from e.g. headlines to stories and speakers to the speech they contribute. In this, we follow our earlier [1] intuition that, in domains like Broadcast News, the material to be summarized exhibits fairly regular patterns from one speech document to another: news broadcasts generally open with a news anchor's introduction of the major news stories to be presented in the broadcast, followed by the actual presentation of those stories by anchor, reporters, and possibly interviewees, and are usually concluded in a fairly conventionalized manner as well. So, we are locating key elements that appear in any broadcast, including different types of speakers (anchor, reporters, interviewees, and soundbite-speakers), anchor signon and signoff, headlines, interviews and soundbites, and news stories themselves. These elements are identified using a combination of acoustic, prosodic, lexical, and structural features obtained from the news transcript and from the original speech. Second, we use similar features to extract portions of news stories to serve as summaries. Thus a newscast can be searched or browsed, to locate stories of interest, and these stories can subsequently be summarized for the user.

The corpus used in the current study is drawn from the TDT2 corpus, a subset of the DARPA HUB-4 Broadcast News corpus. We have annotated 48 hours (96 shows) of CNN shows for named entities, speaker types, anchor sign-on and sign-off, headlines, commercials, interviews, and soundbites. We also make use of the speaker turn, sentence, and story segmentation available in the TDT2 corpus. Finally, we have had a labeler annotate 9 hours (18 programs with a total of 222 stories) of this data for sentences to be included in a summary. Below we provide a brief description of the features we extract from this data for both levels of our summarization approach. We have experimented with a variety of Machine Learning techniques, including Dynamic Bayesian Networks, neural nets, Support Vector Machines, EM methods, and decision trees, on these features to identify key elements for our newscast ’outline’, as well as to identify portions of each story to include in its summary.

We extract three types of features to summarize Broadcast news: structural features, lexical/linguistic features, and acoustic/prosodic features. The structural information we use in our current model follows the approach of [1] in assuming that knowing who the speaker is in a newscast can often tell one what segment of the newscast one is listening to. However, unlike that work, our structural features do not depend upon the explicit identification of speaker type. We take advantage of the fact that more general structural information about the length, position, and overall distribution of speakers’ **turns** — speech segments containing input from a single speaker — can be used directly to select likely candidates for inclusion in a summary of the newscast. The structural information we currently make use of includes the length of each speaker turn, the position of the turn in the overall broadcast, and a calculation of speaker ’type’ based upon the distribution and length of all of a given speakers’ turns in the broadcast. We also use similar information about the previous and subsequent speakers.

The motivation behind using turn position is the observation that key elements such as anchor sig-non and sign-off and headlines typically appear at predictable points in the broadcast. Also, ’important’ information in a broadcast or in a news story appears to come at the begin of its respective unit. Turn length and distribution is an important cue to speaker type (e.g. anchors speaker longer and more frequently than other speakers in a broadcast) and to the usefulness of material to be included in a summary, where interviewee or soundbite speech rarely appears. Information on preceding and subsequent segments provides valuable cues to where a segment itself is located in the broadcast, with typical patterns of anchor/reporter and reporter/interviewee exchanges being common examples. And, when an anchor introduces a news stories, that introductory segment is generally followed by a reporter turn. The anchor’s introductory statement often serves as a short summarize the subsequent story. Modeling the sequence of turns by speaker as well as duration helps us capture such information.

The lexical/linguistic features we use are also useful both for summary extraction and for newscast outlining. To date, we have focussed on simple features, including the presence of noun phrases in general and named entities and their types (person, location, and organization names) in particular, the presence of pronouns, and the length of segments in words. We have found that the presence of multiple named entities of different types is a particularly useful cue to segments to be included in summaries.

Finally, we have experimented with a variety of acoustic/prosodic features, primarily for key element identification – headlines and stories. These include pauses between turns, pitch and energy features, and speaking rate and duration of turns. Segments were examined to extract their f0 range and mean and the difference in these from the prior segment, as well as a ’pitch reset’ feature indicating that the current segment was significantly higher in pitch than previous segments. Several measures of F0 slope were also extracted to find indications of pitch contour fall at the end of segments. We are now including similar features in our story summarization experiments.

Our current results on the identification of headlines achieve 96.9% precision and 63.3% recall, for an F-measure of 76.5% using 10-fold cross validation. Our results for extractive summarization on a small test set currently achieve 75.6% accuracy, with precision 53.7%, reecall 50.6% and an F-measure of 52.1%.

3.2. Summarization of Meetings

4. WHAT IS NEEDED FROM THE SPEECH COMMUNITY

stuff we need from speech community/someone:

segmentation at the sentence, speaker turn, and story level and acoustic/prosodic information used to calculate it (pauses, f0)

named entity extraction.

sentence/clause boundary and disfluency detection (so we can parse the results).

speech act labeling (for summary unit detection?).

confidence scores on words to select ones we are most confident about a la kikuchi et al

phonetic transcription in lattice to get OOV names

5. REFERENCES

[1] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. Identification of speaker role in radio broadcasts. In *Proceedings of AAAI-00*, Austin, 2000.

[2] C. Hori and S. Furui. Advances in automatic speech summarization. In *Proceedings of EUROSPEECH-01*, Aalborg, 2001.

[3] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel. Automatic speech summarization applied to english broadcast news speech. In *Proceedings of ICASSP 2002*, Orlando, 2002.

[4] R. Jin and A. Hauptmann. Automatic title generation for spoken broadcast news. In *Proceedings of ICSLP-00*, Beijing, 2000.

[5] T. Kikuchi, S. Furui, and C. Hori. Two-stage automatic speech summarization by sentence extraction and compaction. In *Proceedings of the IEEE/ISCA Workshop on Spontaneous Speech Processing and Recognition*, pages 207–210, Tokyo, 2003.

[6] S. Maskey and J. Hirschberg. Automatic summarization of broadcast news using structural features. In *Proceedings of EUROSPEECH-03*, Geneva, 2003.