FROM TEXT TO SPEECH SUMMARIZATION

Kathleen McKeown, Julia Hirschberg, Michel Galley and Sameer Maskey

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

ABSTRACT

In this paper, we present approaches used in text summarization, showing how they can be adapted for speech summarization and where they fall short. Informal style and apparent lack of structure in speech mean that the typical approaches used for text summarization must be extended for use with speech. We illustrate how features derived from speech can help determine summary content within two ongoing summarization projects at Columbia University.

1. INTRODUCTION

Text summarization has reached a relatively mature stage; there are well established methods for summarization of a single document and many researchers are working on techniques for summarizing a set of related documents. In this paper, we present approaches used in text summarization, showing how they can be adapted for speech summarization and where they fall short. Given errors resulting from speech recognition and the fact that spoken language is often less formal than written language, the most widely used method for single document text summarization, sentence extraction, cannot be directly applied to speech summarization. However, if systems exploit the additional information that can be derived from the speech signal and from dialog structure, extractive methods can be extended for spoken language and augmented by new methods that focus on extracting particular kinds of information and reformulating it appropriately. We present ongoing work at Columbia on summarization for two different types of spoken sources, broadcast news and meetings.

2. APPROACHES TO SUMMARIZATION

Current summarization systems can be categorized by the type of input that they handle, whether single documents or multiple, and by the approach, whether extractive or abstractive.

To allow summarization in arbitrary domains, most current single document summarization 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. Early approaches used statistical metrics (e.g., word frequencies and key phrases) to identify important sentences (see [14] for papers on the many approaches presented here). 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.

Extractive systems tend to produce summaries with very long sentences; longer sentences score higher on metrics that rate them for importance. Abstractive approaches to single document summarization address this problem by editing the extracted sentences. They 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, the words it contains and their relation to surrounding sentences. 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. Approaches for text compression have used symbolic reduction rules [6], as well as an aligned corpus of documents and their human written summaries to determine which constituents can be reduced [10, 12].

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 [7, 3] producing sets of sentences where each set, or *theme*, contains sentences saying roughly the same thing. Extractive approaches will extract

THIS WORK WAS SUPPORTED IN PART BY NSF GRANT IIS-012196 AND BY DARPA TIDES GRANT NUU01-00-1-8919. ANY OPINIONS, FINDINGS, OR RECOMMENDATIONS ARE THOSE OF THE AUTHORS AND DO NOT NECESSARILY REFLECT THE VIEWS OF THE FUNDING AGENCIES.

one sentence from each set to form the summary. Other multi-document extractive approaches find and extract information about the centroid of the documents [17] or use spreading activation and graph matching to compute similarities and differences between the salient topics of two articles [13].

Only a few researchers have developed abstractive approaches for multi-document summarization. An approach based on information fusion [1] 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 statistical language generation is used to fuse the phrases forming a novel sentence for the summary. Earlier work on multi-document summarization (e.g., [18]) used a symbolic approach, pairing information extraction with language generation. This type of approach produces more of a briefing than a summary. The system 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. Because it generates a summary from structured documents, it can highlight differences as well as similarities.

3. SUMMARIZATION OF SPOKEN LANGUAGE

Speech summarization is a much harder task than text summarization. It is more difficult to identify utterance boundaries; utterances may be fragmentary and may contain disfluencies; and speech recognition introduces errors. Style and lack of explicit formatting mean that the extractive approaches used for text summarization will be more difficult to apply to spoken data. We still need to be able to identify utterances that convey important content, and, given an alignment with an Automatic Speech Recognition (ASR) transcription, we can extract such sentences from the original speech. However, in order to produce a coherent speech summary we need also to develop approaches that can substantially alter the extracted material when we concatenate these segments to form an audio summary. Alternatively, to produce an intelligible text summary of a spoken document, we will need to correct the errors arising from the ASR process and to detect and 'correct' the disfluencies that occur in normal speech. Given these difficulties, summarization of spoken sources has, to date, included single document summarization only.

However, speech summarization also presents opportunities that do not exist for text summarization. Information from the speech signal, such as prosody, can help a system to identify important content and provides good cues to spoken document structure. Information about the speakers can also help determine importance and structure; who is speaking, where the turn falls in relation to other speakers, and how the dialog is structured are important clues. Finally, speech summarization has the option of producing a spoken summary in place of a written summary; in this case, errors in the transcript may not be a problem if the extracted segments are concatenated and replayed, although new issues will need to be addressed such as speaker changes and unnatural changes in energy or pitch. 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 focused on Broadcast News [9, 11, 15]. Such data closely resembles the newswire data that much work in text summarization has concentrated on. Furthermore, there is a large amount of training data available for study, and automatic speech recognition systems to provide transcriptions of reasonable accuracy. However, most current work assumes that such transcripts will be available and of high quality, on which techniques similar to text summarization techniques can then be employed. For example, [11] has used statistical methods to identify words to include in a summary, based upon linguistic features of the transcribed text, while [9] has used lexical extraction methods to hypothesize headlines for news programs. However, such methods are still limited by the quality of the speech transcription itself and this makes the approach of first transcribing into text and then using text-based summarization methods less than successful. To address this, [11] 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 (e.g. n-gram and structured language models), and compressing the output to include only 'important' and well-recognized words.

In our work at Columbia summarizing Broadcast News [15], we have 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 [2]'s 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 locat-

ing 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.

Structural information that we use in our current model follows the approach of [2] 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 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 experiments on summarization of news stories within the broadcast.

3.2. Summarization of Meetings

Meetings are not very similar to written or broadcast news. They involve multi-party conversation with overlapping speakers; the language is informal and utterances tend to be partial, fragmentary, ungrammatical and include many ellipses and pronouns. Furthermore, unlike a news summary, it is not as clear what a meeting summary should include. As a result, extractive summarization alone is not likely to be successful. At Columbia, we are working on meeting summarization as part of a larger project entitled *Mapping Meetings* [16] where the goal is to create methods for effectively recognizing, browsing and visualizing meetings. Our aim for summarization is to produce a high-level record of what happened in the meeting similar to minutes.

Our work to date has focused on methods for identifying important content and for generating the sentences of a summary. Meetings can be long, covering different topics. We developed an approach for segmenting meetings into topics, each of which can be summarized separately. Within each topic, we may find stretches of controversial discussion before a consensus decision is reached. In order to ultimately identify and record issues under discussion, decisions reached and the pros and cons for such decisions, we have developed a method for identifying agreement and disagreement in dialog. We are currently working on methods for combining our work to date with more traditional extraction to identify the important issues and on the development of statistical language generation techniques to compress extracted utterances using various language models, removing from utterances disfluencies and material that is both unimportant and grammatically optional (e.g., prepositional phrases). We will follow this with methods to merge smaller extracted utterances to form the summary.

Our domain-independent topic segmentation algorithm was developed for multi-party speech [4]. It is a feature-based algorithm which combines knowledge about *content* using a text-based segmentation algorithm as one feature and about *form* using linguistic and durational cues about topic shifts extracted from speech. We used features that we identified as strongly correlated with topic changes such as the presence of many speaker overlaps and broad changes in speaker activity distribution. Our work also shows that some features (e.g. silences and cue phrases) that have been used to segment monologue speech preserve their usefulness in multi-party speech. The segmentation algorithm uses automatically trained decision rules to combine the different features. The embedded text-based algorithm builds on lexical cohesion and has performance comparable to state-of-the-art algorithms based on lexical information. A significant error reduction is obtained by combining the speech and text knowledge sources.

Our research on identification of agreement/disagreement [5] is aimed at identifying decisions made as well arguments

made in the meeting, thus providing the basis for generating summary content for each topic segment. Previous work in automatic identification of agreement/disagreement [8] demonstrates that this is a feasible task when various textual, durational, and acoustic features are available. Our work at Columbia builds on this approach and shows that we can get an improvement in accuracy when contextual information is taken into account. The hypothesis is that pragmatic features that center around previous agreement between speakers in the dialog will influence the determination of agreement/disagreement. For example, a speaker who disagrees with another person once in the conversation is more likely to disagree with him again. Our approach first identifies the addressee in each turn based on a set of lexical, durational and structural features that look both forward and backward in the discourse. Second, it combines this knowledge source with information about previous agreements and disagreements to determine the pragmatic orientation of the current utterance. We model context using Bayesian networks that allows capturing of these pragmatic dependencies and get an improvement in accuracy over [8].

4. INTEGRATING TEXT AND SPEECH ANALYSIS

Development of a sophisticated summarization system for spoken language requires further research in both text and speech analysis and provides a fertile testbed for integration of the two approaches. Some of the areas in which more joint research with the speech community would prove valuable include: segmentation of spoken documents at many levels, depending upon genre: utterance, turn, topic or story; extraction of acoustic and prosodic information (pitch, intensity, timing), which may be useful in segmentation but also in identifying 'important' passages to include in a summary; identification of speakers; more accurate named entity extraction from speech; disfluency detection and techniques for 'correcting' disfluent passages; speech act labeling; and access to phoneme lattices and to word level confidence scores from ASR output, to identify out-of-vocabulary proper names and to identify words recognized with higher confidence.

5. REFERENCES

[1] R. Barzilay. *Information Fusion for Multi-Document Summarization: Paraphrasing and Generation*. PhD thesis, Columbia University, 2003.

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

[3] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*, pages 335–336, 1998.

[4] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *Proc. of the 41st ACL*, 2003.

[5] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proc. of the 42nd ACL*, 2004.

[6] G. Grefenstette. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. In *Proceedings of the AAAI Spring Workshop on Intelligent Text Summarization*, pages 111–115, 1998.

[7] V. Hatzivassiloglou, J. Klavans, and E. Eskin. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[8] D. Hillard, M. Ostendorf, and E. Shriberg. Detection of agreement vs. disagreement in meetings: training with unlabeled data. In *Proc. of HLT/NAACL*, 2003.

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

[10] H. Jing and K. McKeown. Cut and paste based summarization. In *Proceedings of the First NAACL*, pages 178–185, Seattle, Washington, 2000.

[11] 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.

[12] K. Knight and D. Marcu. Statistics-based summarization - step one: Sentence compression. In *Proceeding of AAAI-01*, pages 703–710, Austin, Texas, 2001.

[13] I. Mani and E. Bloedorn. Multi-document summarization by graph search and matching. In *Proceedings of AAAI-97*, pages 622–628, Providence, Rhode Island, 1997.

[14] I. Mani and M. T. Maybury, editors. *Advances in Automatic Summarization*. The MIT Press, Cambridge, Massachusetts, 1999.

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

[16] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The meeting project at ICSI. In *Human Language Technologies Conference, San Diego*, 2001.

[17] D. Radev, H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization*, pages 165–172, 2000.

[18] D. R. Radev and K. R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, **24**(3):469–500, September 1998.