Summarization of Spontaneous Conversations

Xiaodan Zhu
Department of Computer Science
10 King’s College Rd.
Toronto, Ontario, Canada
+1 416 946-3922
xzhu@cs.toronto.edu

Gerald Penn
Department of Computer Science
10 King’s College Rd.
Toronto, Ontario, Canada
+1 416 978-7390
gpenn@cs.toronto.edu

ABSTRACT

Spontaneous conversations are an integral element in many CSCW environments. Although speech is often regarded as the most natural and effective way of communication between human beings, speech data are not efficient for quick review. One solution to help people access speech data efficiently in CSCW environments is to conduct speech summarization. Up till now, most speech summarization research has focused on broadcast news; nevertheless summarizing spontaneous conversations is more valuable for CSCW. The task is also more challenging, for example, spontaneous conversations often contain more speech disfluencies, which need to be coped with properly; they are also more vulnerable to speech recognition errors. This demonstration is built to show the prototype of our summarization system. Compared with previous work, our summarizer addresses the problem further in several important respects. First, the system summarizes spontaneous conversations with a wide variety of information/features that have not been explored before, which improve summarization performance according to our experiments. Second, our summarizer handles speech disfluencies, which in all previous work was either not explicitly handled or removed as noise.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: *Abstracting methods, Indexing methods, Linguistic processing.*

General Terms

Performance, Experimentation.

Keywords

Speech summarization, Spontaneous conversations

1. INTRODUCTION

Speech is a natural and efficient way of communication between human beings, so it is an integral element in many CSCW environments, such as videoconferencing. Speech data, however, are not efficient for quick review. To help people access them efficiently, speech summarization can be applied to distill important information and present summaries to users. Speech summarization is a rather new research topic compared with its textual counterpart, but has received increasing interest in the last several years, in domains such as broadcast news stories [1][4]. Compared with broadcast news, spontaneous conversations are more relevant to CSCW, but have received less study in the literature [2][3]. The approaches used in and conclusions drawn from broadcast news or other speech sources do not necessarily fit spontaneous conversations, which are different in many respects: (1) compared with broadcast news, spontaneous conversations are often less well formed linguistically, e.g. They contain more speech disfluencies and false starts; (2) they are also more vulnerable to automatic speech recognition (ASR) errors: word error rates (WERs) of speech recognition are often much higher in spontaneous speech.

Previous work on spontaneous conversation summarization has focused on using textual features, e.g., tf.idf of words [2] and noun senses [3], while speech-related features have not been considered for this type of speech source. Our work addresses the problem further in several important respects. This demonstration is built to show the prototype of our summarization system. Compared with previous work, our summarizer addresses the problem further in several important respects. First, the system summarizes spontaneous conversations with a wide variety of information/features that have not been explored before, which improve summarization performance according to our experiments. Second, our summarizer handles speech disfluencies, which are very common in spontaneous conversations, in a more appropriate way. In all previous work, disfluencies are either not explicitly handled or simply removed as noise.

2. OUR SUMMARIZER

Still in its early stages, research on speech summarization focuses on building extractive, single-document, generic, and surface-level-feature-based summarizers. These extractive summarizers select and present pieces of original speech transcripts or audio segments as summaries, rather than rephrase or rewrite them. The output summary could be textual (transcripts) or spoken (e.g., concatenated audio clips). The pieces to be extracted could correspond to words [1]. The extracts could be utterances, too. Utterance selection is very useful, in that it could be a preliminary stage applied before word extraction (as proposed by Kikuchi et al. [6] in their two-stage summarizer), and with utterance-level extracts, one can play the corresponding audio to users, as with the speech-to-speech summarizer discussed in [7]. The advantage of outputting audio segments rather than transcripts is that it ameliorates the impact of WERs caused by ASR. Therefore, we will focus on utterance-level extraction. The framework of our extractive summarization system is presented figure 1.

CSCW '06, November 4-8, 2006, Banff, Alberta, Canada

2.1 Features extraction

To identify important utterances, we extract and utilize a variety of features: MMR scores, lexical, structural, prosodic features, as well as disfluency features.

- MMR score

  The score calculated with MMR [2] for each utterance.

- Lexical features

  Lexical features include: number of named entities, and utterance length (number of words). The number of named entities include: person-name number, location-name number, organization-name number, and the total number.

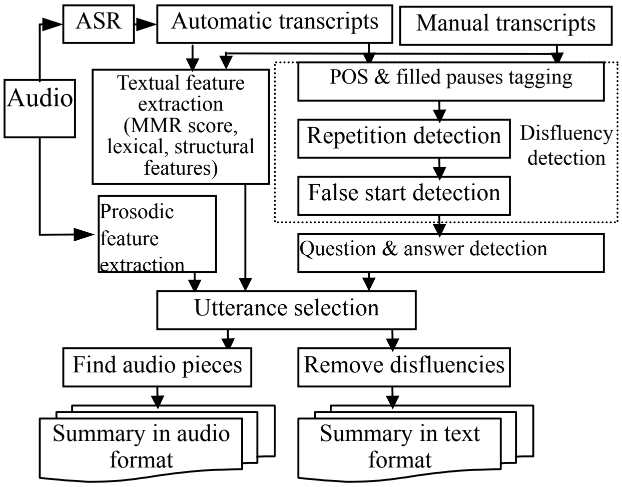

Figure 1. A framework of extractive summarizer for spontaneous conversations

- Structural features

  A value is assigned to indicate whether a given utterance is in the first, middle, or last one-third of the conversation. Another Boolean value is assigned to indicate whether this utterance is adjacent to a speaker turn or not.

- Prosodic features

  Basic prosody includes features such as pitch, energy, speaking rate. They interact with each other and form compound prosody like stress/accentuate, intonation and rhythm. Compound prosody is complicated and difficult to acquire automatically. Same as previous work, we use basic prosody in this paper, the maximum, minimum, average and range of energy, and those of fundamental frequency (f0). These features are calculated on word level and normalized by speakers.

- Disfluency features

  The disfluency features include the number of repetitions, filled-pauses, and the total number of them. Disfluencies adjacent to a speaker turn are ignored here, because they are normally used to coordinate interaction between speakers.

2.2 Disfluency processing

Since disfluencies are very common in spontaneous speech, our summarizer copes with them. Instead of removing them immediately as in [2], disfluency information is fed into the utterance selection module together with other features, because according to [5], disfluencies are not noise and exhibit regularities in a number of dimensions. Later, if the summaries are presented in textual format, we could remove the disfluencies; if the summaries are in audio format, the disfluencies are kept to ensure the naturalness of the summary.

To detect disfluencies, our summarizer follows the approach of [2]. We take as input the manual or automatic transcripts, and use Brill's tagger to assign a part-of-speech (POS) tag to each word. The tag set contains 42 tags, including 38 regular POS tags and four filled-pause tags: CO (empty coordinating conjunctions), DM (lexicalized filled pauses), ET (editing terms), and UH (non-lexicalized filled pauses). Then, repetitions with lengths between 1 and 4 words are detected. Repetitions of greater length are extremely rare and are therefore ignored. Repetitions interrupted by filled-pause words are also detected. False starts are very common in spontaneous speech, too (occurring in 10-15% of utterances). A decision tree (release 8 of C4.5) is used to detect false starts, in the same way as described in [2]. After disfluency detection, question & answer pairs are detected and linked.

2.3 Utterance selection

To obtain a trainable utterance selection module that can utilize and compare rich features, we formulate utterance selection as a standard binary classification problem, and apply two state-of-the-art classifiers, support vector machines (SVM) and logistic regression (LR), to acquire important utterances.

3. DEMOSTRATION

Our demonstration shows the main components of our summarization system, which includes:

- textual feature processing
- prosody processing
- speech disfluency processing
- important utterance selection

We will present summaries in both text and audio formats.

4. REFERENCES

[1] Hori C. and Furui S., 2003. A New Approach to Automatic Speech Summarization IEEE Transactions on Multimedia, Vol. 5, NO. 3, SEPTEMBER 2003, pp. 368-378.

[2] Zechner K., 2001. Automatic Summarization of Spoken Dialogues in Unrestricted Domains. Ph.D. thesis, Carnegie Mellon University, School of Computer Science, Language Technologies Institute, November 2001.

[3] Gurevych I. and Strube M.. 2004. Semantic Similarity Applied to Spoken Dialogue Summarization. In Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, 23-27 August 2004, p.p. 764-770.

[4] Maskey, S.R., Hirschberg, J. "Comparing Lexial, Acoustic/Prosodic, Discourse and Structural Features for Speech Summarization", Eurospeech 2005, Lisbon, Portugal

[5] Shriberg, E.E. (1994). Preliminaries to a Theory of Speech Disfluencies. PhD thesis, University of California at Berkeley.

[6] Kikuchi T., Furui S. and Hori C., 2003. Automatic Speech Summarization Based on Sentence Extraction and Compaction, Proc. ICASSP2003, Hongkong, Vol. I, pp 384-387

[7] Furui, S., Kikuichi T. Shinnaka Y., and Hori C. 2003. Speech-to-speech and speech to text summarization,. First International workshop on Language Understanding and Agents for Real World Interaction, 2003.