UTTERANCE-LEVEL EXTRACTIVE SUMMARIZATION OF OPEN-DOMAIN SPONTANEOUS CONVERSATIONS WITH RICH FEATURES

Xiaodan Zhu and Gerald Penn

Department of Computer Science
Univerisity of Toronto
10 Kings College Rd., Toronto, Canada

ABSTRACT

To identify important utterances from open-domain spontaneous conversations, previous work has focused on using textual features that are extracted from transcripts, e.g., word frequencies and noun senses. In this paper, we summarize spontaneous conversations with features of a wide variety that have not been explored before. Experiments show that the use of speech-related features improves summarization performance. In addition, the effectiveness of individual features is examined and compared.

1. INTRODUCTION

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 or retrieval. One solution to help people access speech data efficiently is through speech summarization. The goal of speech summarization is to distill important information from speech data and to present summaries to users. This is a rather new research topic compared with its textual counterpart, but it has received increasing interest in the last several years, in domains such as voicemail messaging [1], broadcast news stories [2] and spontaneous conversations [3].

Spontaneous conversations are closely related to people's daily life, e.g., telephone conversations. Accordingly, summarizing this type of speech is important. Compared with broadcast news, which has received intensive study [2][5], spontaneous conversations have been less addressed in the literature. Previous work has mainly focused on using textual features, e.g., tf.idf of words [3] and noun senses [4], while speech-related features have not been considered for this type of speech source.

The approaches used in and conclusions drawn from reading news or other speech sources do not necessarily fit spontaneous conversations. Spontaneous conversations are different from reading news in many respects.

First, spontaneous conversations are often much less well formed linguistically, e.g., containing more speech disfluencies. Zechner [3] proposes to detect and remove false starts and speech disfluencies from transcripts. Nevertheless, it is not always necessary to remove disfluencies; for example, original utterances are often more desired to ensure comprehensibility and naturalness if the summaries are to be delivered as excerpts of audio (see section 2). Moreover, disfluencies are not necessary noise; instead, they show regularities in a number of dimensions [7], and correlate with many factors including topic difficulty [6]. Rather than removing them, we explore the effects of disfluencies and repetitions on summarization, which, according to our knowledge, have not been addressed in the literature. Our experiments show that they improve summarization performance.

Second, the distribution of important utterances in spontaneous conversations may differ from that in reading news. For example, important content often appears at the beginning of a news text, but it is usually evenly distributed in conversations. This means that the roles of some features, e.g., structural features, could be different when summarizing these two types of speech. To explore this problem, the effectiveness of individual features is examined and compared. The experiments show that the structural feature is the least effective among all the features used.

Third, conversations often contain discourse clues such as question-answer pairs, which can be utilized to keep the summary coherent. Similar ideas have been proposed recently to summarize online blogs and discussions [11]. This property is not considered in this paper; instead, we focus on applying prosodic and spoken-language features, in addition to textual features, to improve the summarization of spontaneous conversations, and examine the effectiveness of individual features.

2. EXTRACTIVE SUMMARIZATION OF SPONTANEOUS CONVERSATIONS

Still at its early stage, 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 pieces to be extracted could correspond to words. Koumpis [1], for example, extracts important words from transcribed voicemail messages using classification algorithms. Hori & Furui [2] extract words from broadcast news by selecting a path that maximizes a predefined score. Valenza et al. [8] extract N-grams, as well as keywords.

The extracts could be utterances, too. Utterance selection is useful. First, it could be a preliminary stage applied before word extraction, as proposed by Kikuchi et al. [9] in their two-stage summarizer. Second, with utterance-level extracts, one can play the corresponding audio to users, as with the speech-to-speech summarizer discussed in [10]. The advantage of outputting audio segments rather than transcripts is that it avoids the impact of word error rates (WERs) caused by automatic speech recognition (ASR). Therefore, we will focus on utterance-level extraction, which at present appears to be the only way to ensure comprehensibility and naturalness if the summaries are to be delivered as excerpts of audio themselves.

Previous work on utterance extraction from spontaneous conversations mainly focuses on using textual features. Gurevych & Strube [4] develop a shallow knowledge-based approach to extract essential utterances from spontaneous conversation transcripts. To calculate semantic similarity between a given utterance and the conversation, the noun portion of WordNet is used as a knowledge source, with semantic distance between senses computed using Leacock-Chodorow normalized path length. The performance of the system is reported as better than tf.idf based methods. However, the noun senses were manually disambiguated rather than automatically. If the noun senses are instead automatically assigned to be the dominant sense, the summarization performance is worse than tf.idf based maximum marginal relevance (MMR) [13]. Therefore, in our experiments, we use MMR as the baseline.

Zechner [3] applies maximum marginal relevance (MMR) to select utterances for spontaneous conversation transcripts. MMR selects utterances with the following formula:

$$nextUtter = \arg\max_{t_{nr,j}} (\lambda sim_1(t_{nr,j}, query) - (1-\lambda) \max_{t_{r,k}} sim_2(t_{nr,j}, t_{r,k}))$$

MMR ranks utterances by relevance and redundancy. It selects the next unranked utterance into the rank according to two criteria: (1) whether it is more similar to the whole conversation (sim_1 in the formula), and (2) whether it is less similar (sim_2) to the utterances that have so far been selected. Parameter λ linearly combines these two properties. The "query" is a vector of the content words of the spontaneous conversation to be summarized. In [3], MMR is combined with utterance boundary detection, false start detection, repetition filtering, detection of question-answering pairs, and topic segmentation.

3. CLASSIFICATION BASED UTTERANCE EXTRACTION WITH RICH FEATURES

Speech contains more information than textual features, but it is not straightforward to integrate these features into MMR. To utilize these features, we reformulate the utterance selection task as a binary classification problem. An utterance is labeled as either "1" (in-summary) or "0" (not-in-summary). Two state-of-the-art classifiers, support vector machine (SVM) and logistic regression (LR), are used. In this section, we briefly introduce them, and then discuss the features used.

3.1 Classifiers

3.1.1 SVM

A support vector machine (SVM) is a supervised learning technique based on the principle of structural risk minimization. SVM seeks an optimal separating hyperplane, where the margin is maximal. For linearly non-separable samples, SVMs employ the "kernel trick" to implicitly transform the problem to a high-dimensional feature space. The training of SVM solves a quadratic programming problem. In the testing phase, for an input example x, the decision function is:

$$f(x) = sign(\sum_{j=1}^{N_s} a_j y_j K(s_j, x\ ))$$

In our experiment, we use the OSU-SVM package, and use the radial basis kernel.

3.1.2 Logistic Regression

Logistic regression (LR) strives to model the posterior probabilities of the class label with linear functions:

$$\log \frac{p(Y=k \mid X=x)}{p(Y=K \mid X=x)} = \beta_{k0} + \beta_k{}^T x$$

X are feature sets and Y are class labels. For the binary classification that we require in our experiments, the model is especially simple:

$$p(Y=1 \mid X=x) = \frac{\exp(\beta_{10} + \beta_0{}^T x)}{1+\exp(\beta_{10} + \beta_1{}^T x)}$$

$$p(Y=2 \mid X=x) = \frac{1}{1+\exp(\beta_{10} + \beta_1{}^T x)}$$

3.2 Features

The features explored in this paper includes:

- MMR score
  The score calculated with MMR [3] for each utterance.
- Lexicon features
  Lexicon 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. Named entities are annotated automatically with a dictionary.

- 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 pitch, energy, duration, speaking rate and pause. They interact with each other and form compound prosody like stress/accentuate, intonation and rhythm. Compound prosody is complicated and difficult to acquire automatically. In this paper, we use basic prosody, the maximum, minimum, average and range of energy, and those of fundamental frequency (f0).

- Spoken-language features

  The spoken-language features include number of repetitions, disfluencies, and the total number of them. Disfluencies adjacent to a speaker turn are not counted, because they are normally used to coordinate interaction [6] between speakers. Repetitions and disfluencies are acquired in the same way as described in [3].

4. EXPERIMENTAL RESULTS

4.1 Data

The data used for our experiments come from SWITCHBOARD, which is a corpus of telephone conversations, and is widely used in speech-related research. We randomly select 27 conversations, containing around 3660 utterances. The important utterances of each conversation are annotated manually. As with much previous work, we use manual transcripts in this paper.

4.2 Evaluation Metrics

We use two widely used evaluation metrics: f-score and ROUGE score.

4.2.1 F-score

Precision(P)/recall(R) and F-measure are standard evaluation metrics for many NLP tasks. F-score is calculated as $\frac{(\beta+1)*P*R}{(\beta*P+R)}$, where β=1 in our experiments.

4.2.2 ROUGE

ROUGE [12] is a widely used evaluation package for text summarization. It evaluates a summary against gold standards by measuring overlapping units such as n-grams, word sequences, and word pairs.

4.3 Summarization Performance

Ten-fold cross validation is used to obtain the results presented in this section.

4.3.1 F-score

Table-1 shows the f-score of logistic regression (LR) based summarizers, when we generate different lengths of summaries (10-30% of the original utterances).

	10%	15%	20%	25%	30%
(1) MMR	.246	.309	.346	.355	.368
(2) (1)+lexicon	.293	.338	.373	.380	.394
(3) (2)+structure	.334	.366	.400	.409	.404
(4) (3)+acoustic	**.336**	.364	.388	.410	.415
(5) (4)+spoken language	.333	**.376**	**.410**	**.431**	**.422**

Table-1. f-score of LR summarizers using incremental features

Below is the f-score of SVM-based summarizer:

	10%	15%	20%	25%	30%
(1) MMR	.246	.309	.346	.355	.368
(2) (1)+lexicon	.281	.338	.354	.358	.377
(3) (2)+structural	.326	.371	.401	.409	.408
(4) (3)+acoustic	.337	**.380**	.400	.422	.418
(5) (4)+spoken language	**.353**	**.380**	**.416**	**.424**	**.423**

Table 2. f-score of SVM summarizers using incremental features

Both tables show that the performance of the summarizers improved, in general, with more features used. The use of lexicon and structural features outperforms MMR, and the speech-related features, acoustic features and spoken language features, produce additional improvements.

4.3.1 ROUGE

The following tables provide the resulting ROUGE-1 scores:

	10%	15%	20%	25%	30%
(1) MMR	.585	.563	.523	.492	.467
(2) (1)+lexicon	.602	.579	.543	.506	.476
(3) (2)+structure	**.621**	.591	.553	.516	.482
(4) (3)+acoustic	.619	.594	.554	.519	.485
(5) (4)+spoken language	.619	**.600**	**.566**	**.530**	**.492**

Table 3. ROUGE-1 of LR summarizers using incremental features

	10%	15%	20%	25%	30%
(1) MMR	.585	.563	.523	.492	.467
(2) (1)+lexicon	.604	.581	.542	.504	.577
(3) (2)+structure	.617	.600	.563	.523	.490
(4) (3)+acoustic	**.629**	.610	.573	.533	.496
(5) (4)+spoken language	.628	**.611**	**.576**	**.535**	**.502**

Table 4. ROUGE-1 of SVM summarizers using incremental features

The ROUGE-1 scores show similar tendencies to f-scores: the rich features improve summarization performance. Other ROUGE scores like ROUGE-L show the same tendency, but are not presented here due to the space limit.

4.4 Comparison of Features

To study the effectiveness of individual features, the receiver operating characteristic (ROC) curves of these features are drawn in Figure-1 below. ROC is a classic method from signal detection theory and can be used to

clearly compare the effectiveness of features in classification. The larger the area under a curve is, the better the performance of this feature is. To be more exact, the definitions of sensitivity (y-coordinate) and specificity (x-coordinate) are:

$$sensitivity = \frac{TP}{TP + FN} = true\ positive\ rate$$

$$specificity = \frac{TN}{TN + FP} = true\ negtive\ rate$$

where TP, FN, TN and FP are true positive, false negative, true negative, and false positive, respectively.

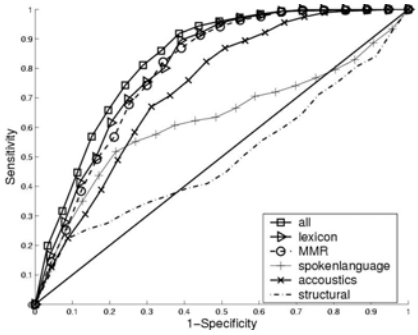

Figure-1. ROC curves for individual features

Lexicon and MMR features are the best feature types when used individually to select utterances, followed by spoken-language and acoustic features. The structural feature is least effective. This may be because of the even distribution of import utterances in spontaneous conversations.

Another perspective is to classify features into "what-is-said" features and "how-it-is-said" features. The former includes MMR and lexicon features; the latter includes structural, acoustic and spoken-language features.

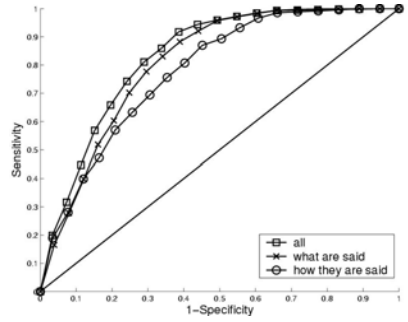

Figure-2. ROC curves for what-are-said and how-they-are-said features

Figure-2 shows that the how-it-is-said features have close performance to the what-is-said feature, i.e., how to say the content provides a comparable amount of information to help identify important utterances.

5. CONCLUSIONS

This paper is concerned with extractive summarization of spontaneous conversations. The task is formulated as a binary classification problem with the use of both text-related and speech-related features. The experiments conducted on SWITCHBOARD show that speech-related features improve summarization performance. The effectiveness of individual features is compared. The MMR and lexicon features are the most effective when used individually, while the structural feature is least effective.

6. REFERENCES

[1] Koumpis K., 2002. Automatic Voicemail Summarisation for Mobile Messaging Ph.D. Thesis in Computer Science, University of Sheffield, UK, 2002.

[2] 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.

[3] 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.

[4] 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.

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

[6] Bortfeld, H., Leon, S.D., Bloom, J.E., Schober, M.F., & Brennan, S.E. 2001. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic Role, and Gender. Language and Speech, 44(2): 123-147

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

[8] Valenza B., Robinson T., Hickey M., Tucker R., 1999. Summarisation of Spoken Audio Through Information Extraction, Proceedings of the ESCA ETRW workshop, 1999.

[9] 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

[10] 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.

[11] Liang Zhou and Eduard Hovy. On the Summarization of Dynamically Introduced Information: Online Discussions and Blogs. To appear in Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, Stanford, CA.

[12] Lin C., 2004. Rouge: a package for automatic evaluation of summaries. In Proceedings of 42st Annual Meeting of the Association for Computational Linguistics, Workshop on Text Summarization Branches Out.

[13] Xiaodan Zhu & Gerald Penn, 2005. Evaluation of Sentence Selection for Speech Summarization. Workshop of Crossing Barriers in Text Summarization, RANLP-2005, Bulgaria.