Comparing the roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization

Xiaodan Zhu **Gerald Penn**

Department of Computer Science, University of Toronto

10 Kings College Rd., Toronto, Canada

{xzhu, gpenn} @cs.toronto.edu

Abstract

This paper is concerned with the summarization of spontaneous conversations. Compared with broadcast news, which has received intensive study, spontaneous conversations have been less addressed in the literature. Previous work has focused on textual features extracted from transcripts. This paper explores and compares the effectiveness of both textual features and speech-related features. The experiments show that these features incrementally improve summarization performance. We also find that speech disfluencies, which have been removed as noise in previous work, help identify important utterances, while the structural feature is less effective than it is in broadcast news.

1 Introduction

Spontaneous conversations are a very important type of speech data. Distilling important information from them has commercial and other importance. Compared with broadcast news, which has received the most intensive studies (Hori and Furui, 2003; Christensen et al. 2004; Maskey and Hirschberg, 2005), spontaneous conversations have been less addressed in the literature.

Spontaneous conversations are different from broadcast news in several aspects: (1) spontaneous conversations are often less well formed linguistically, e.g., containing more speech disfluencies and false starts; (2) the distribution of important utterances in spontaneous conversations could be different from that in broadcast news, e.g., the beginning part of news often contains important information, but in conversations, information may be more evenly distributed; (3) conversations often contain discourse clues, e.g., question-answer pairs and speakers' information, which can be utilized to keep the summary coherent; (4) word error rates (WERs) from speech recognition are usually much higher in spontaneous conversations.

Previous work on spontaneous-conversation summarization has mainly focused on textual features (Zechner, 2001; Gurevych and Strube, 2004), while speech-related features have not been explored for this type of speech source. This paper explores and compares the effectiveness of both textual features and speech-related features. The experiments show that these features incrementally improve summarization performance. We also discuss problems (1) and (2) mentioned above. For (1), Zechner (2001) proposes to detect and remove false starts and speech disfluencies from transcripts, in order to make the text-format summary concise and more readable. Nevertheless, it is not always necessary to remove them. One reason is that 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), in order to avoid the impact of WER. Second, disfluencies are not necessarily noise; instead, they show regularities in a number of dimensions (Shriberg, 1994), and correlate with many factors including topic difficulty (Bortfeld et al, 2001). Rather than removing them, we explore the effects of disfluencies on summarization, which, to our knowledge, has not yet been addressed in the literature. Our experiments show that they improve summarization performance.

To discuss problem (2), we explore and compare both textual features and speech-related features, as they are explored in broadcast news (Maskey and Hirschberg, 2005). The experiments show that the structural feature (e.g. utterance position) is less effective for summarizing spontaneous conversations than it is in broadcast news. MMR

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 197–200,
New York, June 2006.

and lexical features are the best. Speech-related features follow. The structural feature is least effective. We do not discuss problem (3) and (4) in this paper. For problem (3), a similar idea has been proposed to summarize online blogs and discussions. Problem (4) has been partially addressed by (Zechner & Waibel, 2000); but it has not been studied together with acoustic features.

2 Utterance-extraction-based summarization

Still at its early stage, current research on speech summarization targets a less ambitious goal: conducting extractive, single-document, generic, and surface-level-feature-based summarization. The pieces to be extracted could correspond to words (Koumpis, 2002; Hori and Furui, 2003). 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. (2003) 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 Furui et al. (2003). The advantage of outputting audio segments rather than transcripts is that it avoids the impact of WERs caused by automatic speech recognition (ASR). 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 spontaneous conversations mainly focuses on using textual features. Gurevych & Strube (2004) develop a shallow knowledge-based approach. The noun portion of WordNet is used as a knowledge source. The noun senses were manually disambiguated rather than automatically. Zechner (2001) applies maximum marginal relevance (MMR) to select utterances for spontaneous conversation transcripts.

3 Classification based utterance extraction

Spontaneous conversations contain more information than textual features. To utilize these features, we reformulate the utterance selection task as a binary classification problem, an utterance is either labeled as "1" (in-summary) or "0" (not-in-summary). Two state-of-the-art classifiers, support vector machine (SVM) and logistic regression (LR), are used. SVM seeks an optimal separating hyperplane, where the margin is maximal. In our experiments, we use the OSU-SVM package. Logistic regression (LR) is indeed a softmax linear regression, which models the posterior probabilities of the class label with the softmax of linear functions of feature vectors. For the binary classification that we require in our experiments, the model format is simple.

3.1 Features

The features explored in this paper include:

(1) MMR score: the score calculated with MMR (Zechner, 2001) for each utterance.

(2) Lexicon features: number of named entities, and utterance length (number of words). The number of named entities includes: person-name number, location-name number, organization-name number, and the total number. Named entities are annotated automatically with a dictionary.

(3) 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.

(4) Prosodic features: we use basic prosody: the maximum, minimum, average and range of energy, as well as those of fundamental frequency, normalized by speakers. All these features are automatically extracted.

(5) Spoken-language features: the spoken-language features include number of repetitions, filled pauses, and the total number of them. Disfluencies adjacent to a speaker turn are not counted, because they are normally used to coordinate interaction among speakers. Repetitions and pauses are detected in the same way as described in Zechner (2001).

4 Experimental results

4.1 Experiment settings

The data used for our experiments come from SWITCHBOARD. We randomly select 27 conversations, containing around 3660 utterances. The important utterances of each conversation are

manually annotated. We use f-score and the ROUGE score as evaluation metrics. Ten-fold cross validation is applied to obtain the results presented in this section.

4.2 Summarization performance

4.2.1 F-score

Table-1 shows the f-score of logistic regression (LR) based summarizers, under different compression ratios, and with incremental features used.

	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 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.

4.2.2 ROUGE

The following tables provide the 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 the f-scores: the rich features improve summarization performance over the baseline MMR summarizers. Other ROUGE scores like ROUGE-L show the same tendency, but are not presented here due to the space limit.

Both the f-score and ROUGE indicate that, in general, rich features incrementally improve summarization performance.

4.3 Comparison of features

To study the effectiveness of individual features, the receiver operating characteristic (ROC) curves of these features are presented in Figure-1 below. The larger the area under a curve is, the better the performance of this feature is. To be more exact, the definition for the y-coordinate (sensitivity) and the x-coordinate (1-specificity) is:

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

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

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

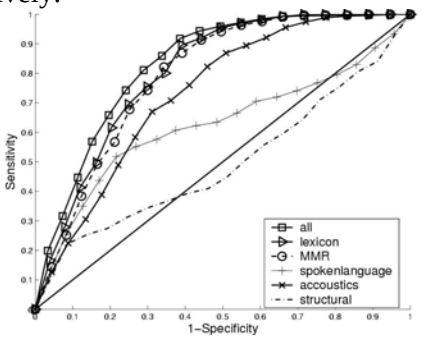

Figure-1. ROC curves for individual features

Lexicon and MMR features are the best two individual features, followed by spoken-language and acoustic features. The structural feature is least effective.

Let us first revisit the problem (2) discussed above in the introduction. The effectiveness of the structural feature is less significant than it is in broadcast news. According to the ROC curves presented in Christensen et al. (2004), the structural feature (utterance position) is one of the best features for summarizing read news stories, and is less effective when news stories contain spontaneous speech. Both their ROC curves cover larger area than the structural feature here in figure 1, that is, the structure feature is less effective for summarizing spontaneous conversation than it is in broadcast news. This reflects, to some extent, that

information is more evenly distributed in spontaneous conversations.

Now let us turn to the role of speech disfluencies, which are very common in spontaneous conversations. Previous work detects and removes disfluencies as noise. Indeed, disfluencies show regularities in a number of dimensions (Shriberg, 1994). They correlate with many factors including the topic difficulty (Bortfeld et al, 2001). Tables 1-4 above show that they improve summarization performance when added upon other features. Figure-1 shows that when used individually, they are better than the structural feature, and also better than acoustic features at the left 1/3 part of the figure, where the summary contains relatively fewer utterances. Disfluencies, e.g., pauses, are often inserted when speakers have word-searching problem, e.g., a problem finding topic-specific keywords:

*Speaker A: with all the **uh** **sulfur** and all that other stuff they're dumping out into the atmosphere.*

The above example is taken from a conversation that discusses pollution. The speaker inserts a filled pause *uh* in front of the word *sulfur*. Pauses are not randomly inserted. To show this, we remove them from transcripts. Section-2 of SWITCHBOARD (about 870 dialogues and 189,000 utterances) is used for this experiment. Then we insert these pauses back randomly, or insert them back at their original places, and compare the difference. For both cases, we consider a window with 4 words after each filled pause. We average the tf.idf scores of the words in each of these windows. Then, for all speaker-inserted pauses, we obtain a set of averaged tf.idf scores. And for all randomly-inserted pauses, we have another set. The mean of the former set (5.79 in table 5) is statistically higher than that of the latter set (5.70 in table 5). We can adjust the window size to 3, 2 and 1, and then get the following table.

Window size		1	2	3	4
Mean of tf.idf score	Insert Randomly	5.69	5.69	5.70	5.70
	Insert by speaker	5.72	5.82	5.81	5.79
Difference is significant? (t-test, $p<0.05$)		Yes	Yes	Yes	Yes

Table 5. Average tf.idf scores of words following filled pauses.

The above table shows that instead of randomly inserting pauses, real speakers insert them in front of words with higher tf.idf scores. This helps explain why disfluencies work.

5 Conclusions

Previous work on summarizing spontaneous conversations has mainly focused on textual features. This paper explores and compares both textual and speech-related features. The experiments show that these features incrementally improve summarization performance. We also find that speech disfluencies, which are removed as noise in previous work, help identify important utterances, while the structural feature is less effective than it is in broadcast news.

6 References

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

Christensen, H., Kolluru, B., Gotoh, Y., Renals, S., 2004. From text summarisation to style-specific summarisation for broadcast news. Proc. ECIR-2004.

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.

Gurevych I. and Strube M. 2004. Semantic Similarity Applied to Spoken Dialogue Summarization. COLING-2004.

Hori C. and Furui S., 2003. A New Approach to Automatic Speech Summarization IEEE Transactions on Multimedia, Vol. 5, NO. 3, September 2003,

Kikuchi T., Furui S. and Hori C., 2003. Automatic Speech Summarization Based on Sentence Extraction and Compaction, Proc. ICASSP-2003.

Koumpis K., 2002. Automatic Voicemail Summarisation for Mobile Messaging Ph.D. Thesis, University of Sheffield, UK, 2002.

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

Shriberg, E.E. (1994). Preliminaries to a Theory of Speech Disfluencies. Ph.D. thesis, University of California at Berkeley.

Zechner K. and Waibel A., 2000. Minimizing word error rate in textual summaries of spoken language. NAACL-2000.

Zechner K., 2001. Automatic Summarization of Spoken Dialogues in Unrestricted Domains. Ph.D. thesis, Carnegie Mellon University, November 2001.