Summarization of Spontaneous Conversations

Xiaodan Zhu & Gerald Penn

Department of Computer Science
University of Toronto, Canada
{xzhu, gpenn}@cs.toronto.edu

Abstract

Most speech summarization research is conducted on broadcast news. In our viewpoint, spontaneous conversations are a more "typical" speech source that distinguishes speech summarization from text summarization, and hence a more appropriate domain for studying speech summarization. For example, spontaneous conversations contain more spoken-language characteristics, e.g. disfluencies and false starts. They are also more vulnerable to ASR errors. Previous research has studied some aspects of this type of data, but this paper addresses the problem further in several important respects. First, we summarize spontaneous conversations with features of a wide variety that have not been explored before. Second, we examine the role of disfluencies in summarization, which in all previous work was either not explicitly handled or removed as noise. Third, we breakdown and analyze the impact of WER on the individual features for summarization.

Index Terms: speech summarization, utterance selection, spontaneous conversations

1. Introduction

The goal of speech summarization is to automatically distill important information from speech data. Most state-of-the-art research is conducted on broadcast news [1][4]. In our viewpoint, spontaneous conversations are a more "typical" speech source that distinguishes speech summarization from text summarization, and hence more appropriate for studying speech summarization. The detailed reasons are: (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 ASR errors: word error rates (WERs) of speech recognition are often much higher in spontaneous speech; (3) spontaneous conversations involve discourse cues, e.g. question-answer pairs and speakers' information, which may be used to keep the summary relevant and coherent.

In addition to its appropriateness, summarizing spontaneous conversations bears great importance too, since they are closely related to people's daily life, e.g. telephone conversations. They are also an integral part of many business activities, e.g. meetings and call-centre custom service. Previous research has addressed some important aspects of this problem [2][3][8]. In this paper, we explore the task further in the following respects: first, we conduct the summarization of spontaneous conversations with features of a wide variety that have not been explored before. Zechner [2] and Gurevych et al [3] use manual transcripts of open-domain dialogues, and textually summarize the transcripts, thus utilizing only textual features to identify important utterances in open-domain dialogues. Murray et al [8] use tf.idf scores of utterances plus prosodic features to select utterances from meeting recordings. The latter also noted that feature-based approaches seem to perform worse than maximum marginal relevance (MMR), although no attempt is made to combine MMR scores as a feature with lexical, structural, prosodic and disfluency features. Unlike [8], our experiments show that rich features can in fact improve summarization performance. We also compare the roles of individual features and find that the structural (utterance-position) features are much less effective in conversation summarization. This is in contradistinction to more "rehearsed" domains such as pure-reading broadcast news and news containing interviews. Secondly, spontaneous conversations are often less well formed, i.e., they contain more speech disfluencies. Zechner [2] detects and removes disfluencies from transcripts, in order to make textual summaries more concise and 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, so as to alleviate the impact of WER caused by ASR. In addition, disfluencies may not actually be noise; on the contrary, they exhibit regularities in a number of dimensions [5]. The possibility therefore remains that disfluencies are actually of use to human speakers in identifying salient information in dialogue, although we are unaware of any psycholinguistic studies that address this claim. Here we instead explore the effects of keeping and using disfluencies on automatic (non-human) summarization, which, according to our knowledge, has also not been addressed in the literature. Our experiments show that they improve summarization performance. Finally, WERs of speech recognition are often high in spontaneous conversations. This paper discusses the impact of this on our utterance-level extractive summarizer. We also breakdown the impact of WER on the individual features for summarization.

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

This research was supported by Avaya Inc., under subproject 3.3b of the NSERC Network for Effective Collaboration Technologies through Advanced Research.

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, using features of a wide variety. The framework of our extractive summarization system is presented below:

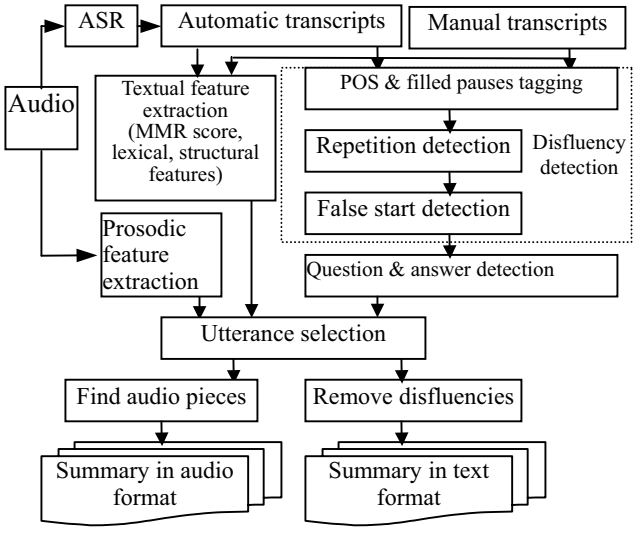

Figure 1. A framework of extractive summarizer for spontaneous conversations

2.1 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. 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. The usefulness of disfluencies is discussed in section 4.

To detect disfluencies, our summarizer follows the approach of [2]. We take as input the manual or automatic transcripts, and use Brill's tagger (Brill, 1995) 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.

2.2 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. 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 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.

2.3 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 have tried several state-of-the-art classifiers, including naive Bayes, LDA, support vector machines (SVM) and logistic regression (LR). In section 4, we present the results with SVM and LR, which achieve the best performance among these in on the SWITCHBOARD dataset.

3. Experimental results

3.1 Experimental settings

The data used for our experiments are the SWITCHBOARD telephone conversations corpus. We randomly select 27 conversations, containing 3665 utterances. The important utterances of each conversation are annotated manually. We use both manual and ASR transcripts, and use ROUGE scores to evaluate the summarizers. ROUGE 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. Ten-fold cross validation is used to obtain the results presented in this section.

3.2 Summarization performance

The following two tables present the results of ROUGE-1 scores for LR and SVM, when we generate different length of

summaries (10-30% of the original utterance number).

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

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

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

Table 2. ROUGE-1 of SVM summarizers using incremental features

Both tables show that the performance of the summarizers improved in general as more features were used. The use of lexical and structural features outperforms Zechner's [2] MMR utterance selection, and speech-related, acoustic and disfluency features produce additional improvements. Our observations differ from those of Murray et al [8]. In that paper, the average and maximum tf.idf scores of utterances, plus prosodic features, were used in Gaussian mixture model and LSA. They observe that their feature-based approaches perform worse than MMR. We speculate that the difference is due to two factors: (1) simply using the average and maximum tf.idf scores might not be appropriate. MMR instead utilizes if.idf to calculate similarities in vector space. Our summarizer directly feed the MMR score as a feature; (2) we use more features in our experiments that are not covered in [8]; all our features are used in addition to the MMR score. Other ROUGE scores like ROUGE-L show the same tendency as the above ROUGE-1 tables.

3.3 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, with the logistic regression classifier. The larger the area under a curve is, the better the performance of this feature is.

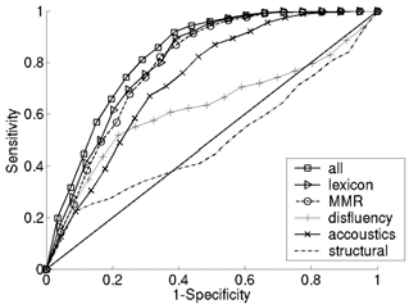

Figure 2. ROC curves for individual features

Lexical features and MMR scores are the best two categories of feature when used individually to select utterances, followed by disfluency and acoustic features. The structural feature (utterance position) is least effective in these spontaneous conversations. We can compared Figure-2 with the ROC curves presented in [9], In [9], the structural feature (utterance position) is one of the best features in summarizing read news stories, although even there it is less effective when news stories contain (more spontaneous) interviews. Both ROC curves in that paper cover a larger area than the structural feature presented in Figure-2, i.e., the structural feature is much more effective in broadcast news. This reflects that information is more evenly distributed in spontaneous conversations.

3.4 Role of disfluencies

In this section, we discuss 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 [5]. Table 1 and 2 above show that disfluencies improve summarization performance when added upon other features. Figure-2 above shows their effectiveness when applied individually to summarization. There, we explore two common categories of disfluencies, filled pauses and repetitions. As mentioned before, we use four types of disfluencies: UH, CO, DM, and ET. A breakdown of effectiveness of them in summarization is shown below:

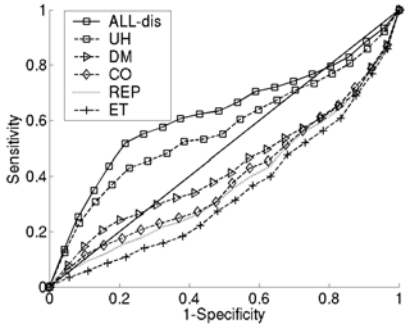

Figure 3. Roles of different types of disfluencies in summarization.

Figure-3 shows that using all these disfluencies together achieve the best summarization performance. Individually, UH words are the most effective type of disfluencies. Actually, they are often inserted when speakers have word-searching problems, e.g. problem of searching for topic-specific keywords or other keywords like named entities:

*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*. UH words are not randomly inserted. To prove this, we remove them from transcripts. Section-2 (filenames begin with *sw2*) 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 into the original places that human speakers did. In both cases, we consider a window with 4 words after each UH word. The average tf.idf scores of these words are then calculated. Therefore, for all UH words inserted by speakers, we obtain an average tf.idf score, and for all randomly-inserted UH words, we obtain another one. We can adjust the window size to 3, 2 and 1, which gives us the following table:

Window size		1	2	3	4
Mean of tf.idf score	Inserted Randomly	5.69	5.69	5.70	5.70
	Inserted by speaker	5.72	5.82	5.81	5.79
Difference is significant?		Yes	Yes	Yes	Yes

Table 3. Average tf.idf scores of words following filled pauses.

As mentioned above, the UH words adjacent to a speaker turn are ignored since they are normally used to coordinate the interaction between speakers. The formulae used to calculate tf and idf in Table 3 are: tf =1+log(raw_tf) and idf = 1+ log (raw_idf), respectively. Before that, stemming is applied. By applying a t-test, we find the difference of tf.idf scores between speaker-inserted and randomly inserted UH words to be significant ($p<0.05$) for each of these window sizes. This means that, instead of inserting UH words randomly, real speakers insert them in front of words with higher tf.idf scores. In addition, UH words may also correlate with certain subtopic structures. Because speakers are more likely to have word-searching problems when they speak certain words for the first time in a conversation; that could exhibit a loose correlation with beginnings of subtopics.

3.5 Impact of ASR word error rates and discourse features

Word error rates (WERs) of speech recognition are usually much higher in spontaneous conversations. In this section, we show their impact on our utterance-extractive summarization. We train ASR models with SWITCHBOARD section-2 data, with the 27 test conversations removed. The word error rate on the test set is 46%. We also train other "pseudo" ASR models with SWITCHBOARD section-2 data, without removing the 27 test conversations. The pseudo-WER on the test data is 39%. We might be able to get less WER by tuning the ASR models or by using more training data, but that is not the focus here. We conduct the summarization on these automatic transcripts and compare the performance with the manual transcripts. The following is the ROUGE-1 score on these transcripts.

WER	10%	15%	20%	25%	30%
0.46	.615	.591	.556	.519	.489
0.39	.615	.591	.557	.526	.491
0	.619	.600	.566	.530	.492

Table 4. ROUGE-1 of LR summarizer under different WERs

Table-4 shows that WERs do not impact the summarization performance significantly. One reason is that the prosodic and structural features are not affected by word errors. Second, although WERs affect MMR-score, disfluency and lexical features, the impact is not very significant. Figure-4 below presents the ROC curves of MMR-score feature and disfluency feature, under different WERs.

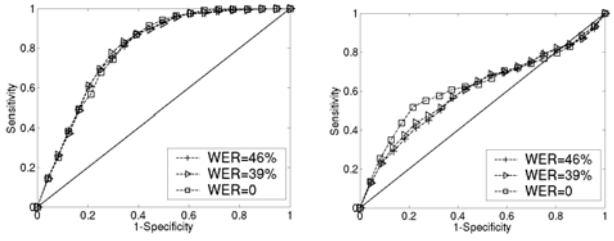

Figure 4. ROC curves for individual features on transcripts with different WERs. The left pane is the effectiveness of MMR scores and the right pane is the effectiveness of disfluencies.

WERs have subtle impact on the MMR-score feature. With analysis, we attribute this to several reasons: (1) keywords are often correctly recognized, although WERs are high. This might be due to the fact that the same topic is discussed in many conversations in Switchboard. In addition, when some keywords are misrecognized (e.g. hat), relevant words (e.g. dress, wear) are also possible to indicate important utterances; (2) a higher WER does not necessarily mean a worse transcript for applications like summarization and classification. Unlike WERs, these applications often concern with bag-of-keywords and do not regard words equally; (3) Utterance length (number of words) is often a latent variable that underlies some other features' roles, e.g., a long utterance often have a higher MMR score than a short utterance, even when the WER changes. We observe that the roles of MMR score in summarization are resistant to WER's change; this is partly because that the utterance length does not change very much when WERs vary.

We also explored the usefulness of discourse information, such as the question & answer labels obtained in Section 2.1, as well as speaker labels. We observed no further improvement of summarization. The reason could be that SWITCHBOARD dialogues have no goal to progress towards—two speakers just chat on a given topic. In a more constructive dialogue, discourse information might be useful; the appropriate use of discourse information deserves further study. Some recent work on summarization of Web blogs or technical discussions shares common characteristics with this problem.

4. Conclusions

In this paper, we summarize spontaneous conversations with rich features, which incrementally improve summarization performance. Utterance position features are found to be much less effective than they are in broadcast news due to the even distribution of information. We also discuss the role of difluencies in summarization. Instead of removing them as noise, we found they improve summarization performance when added upon other features. We also present the summarization performance on transcripts with different WERs.

5. 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.

[8] Gabriel Murray, Steve Renal & Jean Carletta, Extractive Summarization of Meeting Recording, Eurospeech 2005, Lisbon, Portugal

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