A Pilot Study of Opinion Summarization in Conversations

Dong Wang **Yang Liu**
The University of Texas at Dallas
dongwang,yangl@hlt.utdallas.edu

Abstract

This paper presents a pilot study of opinion summarization on conversations. We create a corpus containing extractive and abstractive summaries of speaker's opinion towards a given topic using 88 telephone conversations. We adopt two methods to perform extractive summarization. The first one is a sentence-ranking method that linearly combines scores measured from different aspects including topic relevance, subjectivity, and sentence importance. The second one is a graph-based method, which incorporates topic and sentiment information, as well as additional information about sentence-to-sentence relations extracted based on dialogue structure. Our evaluation results show that both methods significantly outperform the baseline approach that extracts the longest utterances. In particular, we find that incorporating dialogue structure in the graph-based method contributes to the improved system performance.

1 Introduction

Both sentiment analysis (opinion recognition) and summarization have been well studied in recent years in the natural language processing (NLP) community. Most of the previous work on sentiment analysis has been conducted on reviews. Summarization has been applied to different genres, such as news articles, scientific articles, and speech domains including broadcast news, meetings, conversations and lectures. However, opinion summarization has not been explored much. This can be useful for many domains, especially for processing the increasing amount of conversation recordings (telephone conversations, customer service, round-table discussions or interviews in broadcast programs) where we often need to find a person's opinion or attitude, for example, "how does the speaker think about capital punishment and why?". This kind of questions can be treated as a topic-oriented opinion summarization task. Opinion summarization was run as a pilot task in Text Analysis Conference (TAC) in 2008. The task was to produce summaries of opinions on specified targets from a set of blog documents. In this study, we investigate this problem using spontaneous conversations. The problem is defined as, given a conversation and a topic, a summarization system needs to generate a summary of the speaker's opinion towards the topic.

This task is built upon opinion recognition and topic or query based summarization. However, this problem is challenging in that: (a) Summarization in spontaneous speech is more difficult than well structured text (Mckeown et al., 2005), because speech is always less organized and has recognition errors when using speech recognition output; (b) Sentiment analysis in dialogues is also much harder because of the genre difference compared to other domains like product reviews or news resources, as reported in (Raaijmakers et al., 2008); (c) In conversational speech, information density is low and there are often off topic discussions, therefore presenting a need to identify utterances that are relevant to the topic.

In this paper we perform an exploratory study on opinion summarization in conversations. We compare two unsupervised methods that have been

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 331–339,
Portland, Oregon, June 19-24, 2011.

widely used in extractive summarization: sentence-ranking and graph-based methods. Our system attempts to incorporate more information about topic relevancy and sentiment scores. Furthermore, in the graph-based method, we propose to better incorporate the dialogue structure information in the graph in order to select salient summary utterances. We have created a corpus of reasonable size in this study. Our experimental results show that both methods achieve better results compared to the baseline.

The rest of this paper is organized as follows. Section 2 briefly discusses related work. Section 3 describes the corpus and annotation scheme we used. We explain our opinion-oriented conversation summarization system in Section 4 and present experimental results and analysis in Section 5. Section 6 concludes the paper.

2 Related Work

Research in document summarization has been well established over the past decades. Many tasks have been defined such as single-document summarization, multi-document summarization, and query-based summarization. Previous studies have used various domains, including news articles, scientific articles, web documents, reviews. Recently there is an increasing research interest in speech summarization, such as conversational telephone speech (Zhu and Penn, 2006; Zechner, 2002), broadcast news (Maskey and Hirschberg, 2005; Lin et al., 2009), lectures (Zhang et al., 2007; Furui et al., 2004), meetings (Murray et al., 2005; Xie and Liu, 2010), voice mails (Koumpis and Renals, 2005). In general speech domains seem to be more difficult than well written text for summarization. In previous work, unsupervised methods like Maximal Marginal Relevance (MMR), Latent Semantic Analysis (LSA), and supervised methods that cast the extraction problem as a binary classification task have been adopted. Prior research has also explored using speech specific information, including prosodic features, dialog structure, and speech recognition confidence.

In order to provide a summary over opinions, we need to find out which utterances in the conversation contain opinion. Most previous work in sentiment analysis has focused on reviews (Pang and Lee, 2004; Popescu and Etzioni, 2005; Ng et al., 2006) and news resources (Wiebe and Riloff, 2005). Many kinds of features are explored, such as lexical features (unigram, bigram and trigram), part-of-speech tags, dependency relations. Most of prior work used classification methods such as naive Bayes or SVMs to perform the polarity classification or opinion detection. Only a handful studies have used conversational speech for opinion recognition (Murray and Carenini, 2009; Raaijmakers et al., 2008), in which some domain-specific features are utilized such as structural features and prosodic features.

Our work is also related to question answering (QA), especially opinion question answering. (Stoyanov et al., 2005) applies a subjectivity filter based on traditional QA systems to generate opinionated answers. (Balahur et al., 2010) answers some specific opinion questions like "Why do people criticize Richard Branson?" by retrieving candidate sentences using traditional QA methods and selecting the ones with the same polarity as the question. Our work is different in that we are not going to answer specific opinion questions, instead, we provide a summary on the speaker's opinion towards a given topic.

There exists some work on opinion summarization. For example, (Hu and Liu, 2004; Nishikawa et al., 2010) have explored opinion summarization in review domain, and (Paul et al., 2010) summarizes contrastive viewpoints in opinionated text. However, opinion summarization in spontaneous conversation is seldom studied.

3 Corpus Creation

Though there are many annotated data sets for the research of speech summarization and sentiment analysis, there is no corpus available for opinion summarization on spontaneous speech. Thus for this study, we create a new pilot data set using a subset of the Switchboard corpus (Godfrey and Holliman, 1997).[1] These are conversational telephone speech between two strangers that were assigned a topic to talk about for around 5 minutes. They were told to find the opinions of the other person. There are 70 topics in total. From the Switchboard cor-

[1] Please contact the authors to obtain the data.

pus, we selected 88 conversations from 6 topics for this study. Table 1 lists the number of conversations in each topic, their average length (measured in the unit of dialogue acts (DA)) and standard deviation of length.

topic	#Conv.	avg len	stdev
space flight and exploration	6	165.5	71.40
capital punishment	24		
gun control	15		
universal health insurance	9		
drug testing	12		
universal public service	22		

Table 1: Corpus statistics: topic description, number of conversations in each topic, average length (number of dialog acts), and standard deviation.

We recruited 3 annotators that are all undergraduate computer science students. From the 88 conversations, we selected 18 (3 from each topic) and let all three annotators label them in order to study inter-annotator agreement. The rest of the conversations has only one annotation.

The annotators have access to both conversation transcripts and audio files. For each conversation, the annotator writes an abstractive summary of up to 100 words for each speaker about his/her opinion or attitude on the given topic. They were told to use the words in the original transcripts if possible. Then the annotator selects up to 15 DAs (no minimum limit) in the transcripts for each speaker, from which their abstractive summary is derived. The selected DAs are used as the human generated extractive summary. In addition, the annotator is asked to select an overall opinion towards the topic for each speaker among five categories: strongly support, somewhat support, neutral, somewhat against, strongly against. Therefore for each conversation, we have an abstractive summary, an extractive summary, and an overall opinion for each speaker. The following shows an example of such annotation for speaker B in a dialogue about "capital punishment":

[**Extractive Summary**]

I think I've seen some statistics that say that, uh, it's more expensive to kill somebody than to keep them in prison for life.

committing them mostly is, you know, either crimes of passion or at the moment

or they think they're not going to get caught

but you also have to think whether it's worthwhile on the individual basis, for example, someone like, uh, jeffrey dahlmer,

by putting him in prison for life, there is still a possibility that he will get out again.

I don't think he could ever redeem himself,

but if you look at who gets accused and who are the ones who actually get executed, it's very racially related – and ethnically related

[**Abstractive Summary**]

B is against capital punishment except under certain circumstances. B finds that crimes deserving of capital punishment are "crimes of the moment" and as a result feels that capital punishment is not an effective deterrent. however, B also recognizes that on an individual basis some criminals can never "redeem" themselves.

[**Overall Opinion**]

Somewhat against

Table 2 shows the compression ratio of the extractive summaries and abstractive summaries as well as their standard deviation. Because in conversations, utterance length varies a lot, we use words as units when calculating the compression ratio.

	avg ratio	stdev
extractive summaries	0.26	0.13
abstractive summaries	0.13	0.06

Table 2: Compression ratio and standard deviation of extractive and abstractive summaries.

We measured the inter-annotator agreement among the three annotators for the 18 conversations (each has two speakers, thus 36 "documents" in total). Results are shown in Table 3. For the extractive or abstractive summaries, we use ROUGE scores (Lin, 2004), a metric used to evaluate automatic summarization performance, to measure the pairwise agreement of summaries from different annotators. ROUGE F-scores are shown in the table for different matches, unigram (R-1), bigram (R-2), and longest subsequence (R-L). For the overall opinion category, since it is a multiclass label (not binary decision), we use Krippendorff's α coefficient to measure human agreement, and the difference function for interval data: $\delta^2_{ck} = (c-k)^2$ (where c, k are the interval values, on a scale of 1 to 5 corresponding to the five categories for the overall opinion).

We notice that the inter-annotator agreement for extractive summaries is comparable to other speech

extractive summaries	R-1	0.61
	R-2	0.52
	R-L	0.61
abstractive summaries	R-1	0.32
	R-2	0.13
	R-L	0.25
overall opinion	$\alpha = 0.79$	

Table 3: Inter-annotator agreement for extractive and abstractive summaries, and overall opinion.

summary annotation (Liu and Liu, 2008). The agreement on abstractive summaries is much lower than extractive summaries, which is as expected. Even for the same opinion or sentence, annotators use different words in the abstractive summaries. The agreement for the overall opinion annotation is similar to other opinion/emotion studies (Wilson, 2008b), but slightly lower than the level recommended by Krippendorff for reliable data ($\alpha = 0.8$) (Hayes and Krippendorff, 2007), which shows it is even difficult for humans to determine what opinion a person holds (support or against something). Often human annotators have different interpretations about the same sentence, and a speaker's opinion/attitude is sometimes ambiguous. Therefore this also demonstrates that it is more appropriate to provide a summary rather than a simple opinion category to answer questions about a person's opinion towards something.

4 Opinion Summarization Methods

Automatic summarization can be divided into extractive summarization and abstractive summarization. Extractive summarization selects sentences from the original documents to form a summary; whereas abstractive summarization requires generation of new sentences that represent the most salient content in the original documents like humans do. Often extractive summarization is used as the first step to generate abstractive summary.

As a pilot study for the problem of opinion summarization in conversations, we treat this problem as an extractive summarization task. This section describes two approaches we have explored in generating extractive summaries. The first one is a sentence-ranking method, in which we measure the salience of each sentence according to a linear combination of scores from several dimensions. The second one is a graph-based method, which incorporates the dialogue structure in ranking. We choose to investigate these two methods since they have been widely used in text and speech summarization, and perform competitively. In addition, they do not require a large labeled data set for modeling training, as needed in some classification or feature based summarization approaches.

4.1 Sentence Ranking

In this method, we use Equation 1 to assign a score to each DA s, and select the most highly ranked ones until the length constriction is satisfied.

$$\begin{aligned} score(s) &= \lambda_{sim} sim(s, D) + \lambda_{rel} REL(s, topic) \\ &\quad + \lambda_{sent} sentiment(s) + \lambda_{len} length(s) \\ \sum_i \lambda_i &= 1 \end{aligned} \quad (1)$$

- $sim(s, D)$ is the cosine similarity between DA s and all the utterances in the dialogue from the same speaker, D. It measures the relevancy of s to the entire dialogue from the target speaker. This score is used to represent the salience of the DA. It has been shown to be an important indicator in summarization for various domains. For cosine similarity measure, we use TF*IDF (term frequency, inverse document frequency) term weighting. The IDF values are obtained using the entire Switchboard corpus, treating each conversation as a document.

- $REL(s, topic)$ measures the topic relevance of DA s. It is the sum of the topic relevance of all the words in the DA. We only consider the content words for this measure. They are identified using TreeTagger toolkit.[2] To measure the relevance of a word to a topic, we use Pairwise Mutual Information (PMI):

$$PMI(w, topic) = log_2 \frac{p(w\&topic)}{p(w)p(topic)} \quad (2)$$

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

where all the statistics are collected from the Switchboard corpus: $p(w\&topic)$ denotes the probability that word w appears in a dialogue of topic t, and $p(w)$ is the probability of w appearing in a dialogue of any topic. Since our goal is to rank DAs in the same dialog, and the topic is the same for all the DAs, we drop $p(topic)$ when calculating PMI scores. Because the value of $PMI(w, topic)$ is negative, we transform it into a positive one (denoted by $PMI^+(w, topic)$) by adding the absolute value of the minimum value. The final relevance score of each sentence is normalized to $[0, 1]$ using linear normalization:

$$\begin{aligned} REL_{orig}(s, topic) &= \sum_{w \in s} PMI^+(w, topic) \\ REL(s, topic) &= \frac{REL_{orig}(s, topic) - Min}{Max - Min} \end{aligned}$$

- $sentiment(s)$ indicates the probability that utterance s contains opinion. To obtain this, we trained a maximum entropy classifier with a bag-of-words model using a combination of data sets from several domains, including movie data (Pang and Lee, 2004), news articles from MPQA corpus (Wilson and Wiebe, 2003), and meeting transcripts from AMI corpus (Wilson, 2008a). Each sentence (or DA) in these corpora is annotated as "subjective" or "objective". We use each utterance's probability of being "subjective" predicted by the classifier as its sentiment score.

- $length(s)$ is the length of the utterance. This score can effectively penalize the short sentences which typically do not contain much important content, especially the backchannels that appear frequently in dialogues. We also perform linear normalization such that the final value lies in $[0, 1]$.

4.2 Graph-based Summarization

Graph-based methods have been widely used in document summarization. In this approach, a document is modeled as an adjacency matrix, where each node represents a sentence, and the weight of the edge between each pair of sentences is their similarity (cosine similarity is typically used). An iterative process is used until the scores for the nodes converge. Previous studies (Erkan and Radev, 2004) showed that this method can effectively extract important sentences from documents. The basic framework we use in this study is similar to the query-based graph summarization system in (Zhao et al., 2009). We also consider sentiment and topic relevance information, and propose to incorporate information obtained from dialog structure in this framework. The score for a DA s is based on its content similarity with all other DAs in the dialogue, the connection with other DAs based on the dialogue structure, the topic relevance, and its subjectivity, that is:

$$\begin{aligned} score(s) &= \lambda_{sim} \sum_{v \in C} \frac{sim(s, v)}{\sum_{z \in C} sim(z, v)} score(v) \\ &\quad + \lambda_{rel} \frac{REL(s, topic)}{\sum_{z \in C} REL(z, topic)} \\ &\quad + \lambda_{sent} \frac{sentiment(s)}{\sum_{z \in C} sentiment(z)} \\ &\quad + \lambda_{adj} \sum_{v \in C} \frac{ADJ(s, v)}{\sum_{z \in C} ADJ(z, v)} score(v) \\ \sum_i \lambda_i &= 1 \end{aligned} \quad (3)$$

where C is the set of all DAs in the dialogue; $REL(s, topic)$ and $sentiment(s)$ are the same as those in the above sentence ranking method; $sim(s, v)$ is the cosine similarity between two DAs s and v. In addition to the standard connection between two DAs with an edge weight $sim(s, v)$, we introduce new connections $ADJ(s, v)$ to model dialog structure. It is a directed edge from s to v, defined as follows:

- If s and v are from the same speaker and within the same turn, there is an edge from s to v and an edge from v to s with weight $1/dis(s, v)$ ($ADJ(s, v) = ADJ(v, s) = 1/dis(s, v)$), where $dis(s, v)$ is the distance between s and v, measured based on their DA indices. This way the DAs in the same turn can reinforce each other. For example, if we consider that

one DA is important, then the other DAs in the same turn are also important.

- If s and v are from the same speaker, and separated only by one DA from another speaker with length less than 3 words (usually backchannel), there is an edge from s to v as well as an edge from v to s with weight 1 ($ADJ(s, v) = ADJ(v, s) = 1$).

- If s and v form a question-answer pair from two speakers, then there is an edge from question s to answer v with weight 1 ($ADJ(s, v) = 1$). We use a simple rule-based method to determine question-answer pairs — sentence s has question marks or contains "wh-word" (i.e., "what, how, why"), and sentence v is the immediately following one. The motivation for adding this connection is, if the score of a question sentence is high, then the answer's score is also boosted.

- If s and v form an agreement or disagreement pair, then there is an edge from v to s with weight 1 ($ADJ(v, s) = 1$). This is also determined by simple rules: sentence v contains the word "agree" or "disagree", s is the previous sentence, and from a different speaker. The reason for adding this is similar to the above question-answer pairs.

- If there are multiple edges generated from the above steps between two nodes, then we use the highest weight.

Since we are using a directed graph for the sentence connections to model dialog structure, the resulting adjacency matrix is asymmetric. This is different from the widely used graph methods for summarization. Also note that in the first sentence ranking method or the basic graph methods, summarization is conducted for each speaker separately. Utterances from one speaker have no influence on the summary decision for the other speaker. Here in our proposed graph-based method, we introduce connections between the two speakers, so that the adjacency pairs between them can be utilized to extract salient utterances.

5 Experiments

5.1 Experimental Setup

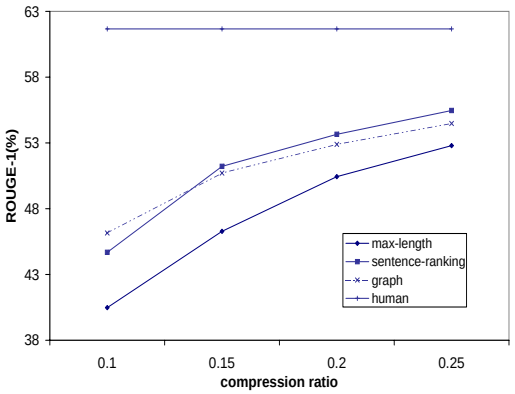
The 18 conversations annotated by all 3 annotators are used as test set, and the rest of 70 conversations are used as development set to tune the parameters (determining the best combination weights). In preprocessing we applied word stemming. We perform extractive summarization using different word compression ratios (ranging from 10% to 25%). We use human annotated dialogue acts (DA) as the extraction units. The system-generated summaries are compared to human annotated extractive and abstractive summaries. We use ROUGE as the evaluation metrics for summarization performance.

We compare our methods to two systems. The first one is a baseline system, where we select the longest utterances for each speaker. This has been shown to be a relatively strong baseline for speech summarization (Gillick et al., 2009). The second one is human performance. We treat each annotator's extractive summary as a system summary, and compare to the other two annotators' extractive and abstractive summaries. This can be considered as the upper bound of our system performance.

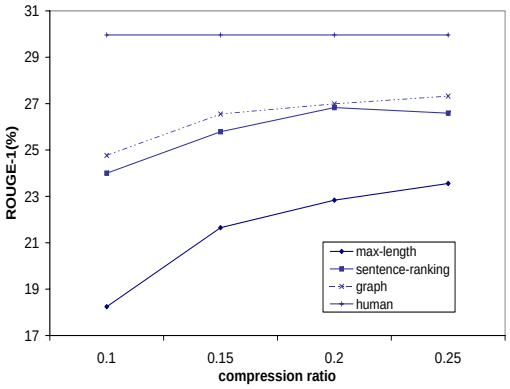
5.2 Results

From the development set, we used the grid search method to obtain the best combination weights for the two summarization methods. In the sentence-ranking method, the best parameters found on the development set are $\lambda_{sim} = 0, \lambda_{rel} = 0.3, \lambda_{sent} = 0.3, \lambda_{len} = 0.4$. It is surprising to see that the similarity score is not useful for this task. The possible reason is, in Switchboard conversations, what people talk about is diverse and in many cases only topic words (except stopwords) appear more than once. In addition, REL score is already able to catch the topic relevancy of the sentence. Thus, the similarity score is redundant here.

In the graph-based method, the best parameters are $\lambda_{sim} = 0, \lambda_{adj} = 0.3, \lambda_{rel} = 0.4, \lambda_{sent} = 0.3$. The similarity between each pair of utterances is also not useful, which can be explained with similar reasons as in the sentence-ranking method. This is different from graph-based summarization systems for text domains. A similar finding has also been shown in (Garg et al., 2009), where similarity be-



(a) compare to reference extractive summary



(b) compare to reference abstractive summary

Figure 1: ROUGE-1 F-scores compared to extractive and abstractive reference summaries for different systems: max-length, sentence-ranking method, graph-based method, and human performance.

tween utterances does not perform well in conversation summarization.

Figure 1 shows the ROUGE-1 F-scores comparing to human extractive and abstractive summaries for different compression ratios. Similar patterns are observed for other ROUGE scores such as ROUGE-2 or ROUGE-L, therefore they are not shown here. Both methods improve significantly over the baseline approach. There is relatively less improvement using a higher compression ratio, compared to a lower one. This is reasonable because when the compression ratio is low, the most salient utterances are not necessarily the longest ones, thus using more information sources helps better identify important sentences; but when the compression ratio is higher, longer utterances are more likely to be selected since they contain more content.

There is no significant difference between the two methods. When compared to extractive reference summaries, sentence-ranking is slightly better except for the compression ratio of 0.1. When compared to abstractive reference summaries, the graph-based method is slightly better. The two systems share the same topic relevance score (REL) and sentiment score, but the sentence-ranking method prefers longer DAs and the graph-based method prefers DAs that are emphasized by the ADJ matrix, such as the DA in the middle of a cluster of utterances from the same speaker, the answer to a question, etc.

5.3 Analysis

To analyze the effect of dialogue structure we introduce in the graph-based summarization method, we compare two configurations: $\lambda_{adj} = 0$ (only using REL score and sentiment score in ranking) and $\lambda_{adj} = 0.3$. We generate summaries using these two setups and compare with human selected sentences. Table 4 shows the number of false positive instances (selected by system but not by human) and false negative ones (selected by human but not by system). We use all three annotators' annotation as reference, and consider an utterance as positive if one annotator selects it. This results in a large number of reference summary DAs (because of low human agreement), and thus the number of false negatives in the system output is very high. As expected, a smaller compression ratio (fewer selected DAs in the system output) yields a higher false negative rate and a lower false positive rate. From the results, we can see that generally adding adjacency matrix information is able to reduce both types of errors except when the compression ratio is 0.15.

The following shows an example, where the third DA is selected by the system with $\lambda_{adj} = 0.3$, but not by $\lambda_{adj} = 0$. This is partly because the weight of the second DA is enhanced by the the question-

	$\lambda_{adj} = 0$		$\lambda_{adj} = 0.3$	
ratio	FP	FN	FP	FN
0.1	37	588	33	581
0.15	60	542	61	546
0.2	100	516	90	511
0.25	137	489	131	482

Table 4: The number of false positive (FP) and false negative (FN) instances using the graph-based method with $\lambda_{adj} = 0$ and $\lambda_{adj} = 0.3$ for different compression ratios.

answer pair (the first and the second DA), and thus subsequently boosting the score of the third DA.

A: Well what do you think?

B: Well, I don't know, I'm thinking about from one to ten what my no would be.

B: It would probably be somewhere closer to, uh, less control because I don't see, -

We also examined the system output and human annotation and found some reasons for the system errors:

(a) Topic relevance measure. We use the statistics from the Switchboard corpus to measure the relevance of each word to a given topic (PMI score), therefore only when people use the same word in different conversations of the topic, the PMI score of this word and the topic is high. However, since the size of the corpus is small, some topics only contain a few conversations, and some words only appear in one conversation even though they are topic-relevant. Therefore the current PMI measure cannot properly measure a word's and a sentence's topic relevance. This problem leads to many false negative errors (relevant sentences are not captured by our system).

(b) Extraction units. We used DA segments as units for extractive summarization, which can be problematic. In conversational speech, sometimes a DA segment is not a complete sentence because of overlaps and interruptions. We notice that annotators tend to select consecutive DAs that constitute a complete sentence, however, since each individual DA is not quite meaningful by itself, they are often not selected by the system. The following segment is extracted from a dialogue about "universal health insurance". The two DAs from speaker B are not selected by our system but selected by human annotators, causing false negative errors.

B: and it just can devastate –
A: and your constantly, -
B: – your budget, you know.

6 Conclusion and Future Work

This paper investigates two unsupervised methods in opinion summarization on spontaneous conversations by incorporating topic score and sentiment score in existing summarization techniques. In the sentence-ranking method, we linearly combine several scores in different aspects to select sentences with the highest scores. In the graph-based method, we use an adjacency matrix to model the dialogue structure and utilize it to find salient utterances in conversations. Our experiments show that both methods are able to improve the baseline approach, and we find that the cosine similarity between utterances or between an utterance and the whole document is not as useful as in other document summarization tasks.

In future work, we will address some issues identified from our error analysis. First, we will investigate ways to represent a sentence's topic relevance. Second, we will evaluate using other extraction units, such as applying preprocessing to remove disfluencies and concatenate incomplete sentence segments together. In addition, it would be interesting to test our system on speech recognition output and automatically generated DA boundaries to see how robust it is.

7 Acknowledgments

The authors thank Julia Hirschberg and Ani Nenkova for useful discussions. This research is supported by NSF awards CNS-1059226 and IIS-0939966.

References

Alexandra Balahur, Ester Boldrini, Andrés Montoyo, and Patricio Martínez-Barco. 2010. *Going beyond traditional QA systems: challenges and keys in opinion question answering*. In *Proceedings of COLING*.

Günes Erkan and Dragomir R. Radev. 2004. *LexRank: graph-based lexical centrality as salience in text summarization*. *Journal of Artificial Intelligence Research*.

Sadaoki Furui, Tomonori Kikuchi, Yousuke Shinnaka, and Chior i Hori. 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Audio, Speech & Language Processing*, 12(4):401–408.

Nikhil Garg, Benoit Favre, Korbinian Reidhammer, and Dilek Hakkani Tür. 2009. *ClusterRank: a graph based method for meeting summarization*. In *Proceedings of Interspeech*.

Dan Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. 2009. *A global optimization framework for meeting summarization*. In *Proceedings of ICASSP*.

John J. Godfrey and Edward Holliman. 1997. *Switchboard-1 Release 2*. In *Linguistic Data Consortium, Philadelphia*.

Andrew Hayes and Klaus Krippendorff. 2007. *Answering the call for a standard reliability measure for coding data*. *Journal of Communication Methods and Measures*, 1:77–89.

Minqing Hu and Bing Liu. 2004. *Mining and summarizing customer reviews*. In *Proceedings of ACM SIGKDD*.

Konstantinos Koumpis and Steve Renals. 2005. *Automatic summarization of voicemail messages using lexical and prosodic features*. *ACM - Transactions on Speech and Language Processing*.

Shih Hsiang Lin, Berlin Chen, and Hsin min Wang. 2009. A comparative study of probabilistic ranking models for chinese spoken document summarization. *ACM Transactions on Asian Language Information Processing*, 8(1).

Chin-Yew Lin. 2004. *ROUGE: a package for automatic evaluation of summaries*. In *Proceedings of ACL workshop on Text Summarization Branches Out*.

Fei Liu and Yang Liu. 2008. *What are meeting summaries? An analysis of human extractive summaries in meeting corpus*. In *Proceedings of SIGDial*.

Sameer Maskey and Julia Hirschberg. 2005. *Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization*. In *Proceedings of Interspeech*.

Kathleen Mckeown, Julia Hirschberg, Michel Galley, and Sameer Maskey. 2005. *From text to speech summarization*. In *Proceedings of ICASSP*.

Gabriel Murray and Giuseppe Carenini. 2009. *Detecting subjectivity in multiparty speech*. In *Proceedings of Interspeech*.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. *Extractive summarization of meeting recordings*. In *Proceedings of EUROSPEECH*.

Vincent Ng, Sajib Dasgupta, and S.M.Niaz Arifin. 2006. *Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews*. In *Proceedings of the COLING/ACL*.

Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo, and Genichiro Kikui. 2010. *Opinion summarization with integer linear programming formulation for sentence extraction and ordering*. In *Proceedings of COLING*.

Bo Pang and Lilian Lee. 2004. *A sentiment education: sentiment analysis using subjectivity summarization based on minimum cuts*. In *Proceedings of ACL*.

Michael Paul, ChengXiang Zhai, and Roxana Girju. 2010. *Summarizing contrastive viewpoints in opinionated text*. In *Proceedings of EMNLP*.

Ana-Maria Popescu and Oren Etzioni. 2005. *Extracting product features and opinions from reviews*. In *Proceedings of HLT-EMNLP*.

Stephan Raaijmakers, Khiet Truong, and Theresa Wilson. 2008. *Multimodal subjectivity analysis of multiparty conversation*. In *Proceedings of EMNLP*.

Veselin Stoyanov, Claire Cardie, and Janyce Wiebe. 2005. *Multi-perspective question answering using the OpQA corpus*. In *Proceedings of EMNLP/HLT*.

Janyce Wiebe and Ellen Riloff. 2005. *Creating subjective and objective sentence classifiers from unannotated texts*. In *Proceedings of CICLing*.

Theresa Wilson and Janyce Wiebe. 2003. *Annotating opinions in the world press*. In *Proceedings of SIGDial*.

Theresa Wilson. 2008a. *Annotating subjective content in meetings*. In *Proceedings of LREC*.

Theresa Wilson. 2008b. Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states. Ph.D. thesis, University of Pittsburgh.

Shasha Xie and Yang Liu. 2010. Improving supervised learning for meeting summarization using sampling and regression. *Computer Speech and Language*, 24:495–514.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in dive rse genres. *Computational Linguistics*, 28:447–485.

Justin Jian Zhang, Ho Yin Chan, and Pascale Fung. 2007. *Improving lecture speech summarization using rhetorical information*. In *Proceedings of Biannual IEEE Workshop on ASRU*.

Lin Zhao, Lide Wu, and Xuanjing Huang. 2009. *Using query expansion in graph-based approach for query-focused multi-document summarization*. *Journal of Information Processing and Management*.

Xiaodan Zhu and Gerald Penn. 2006. *Summarization of spontaneous conversations*. In *Proceedings of Interspeech*.