Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 61–82

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Opinion summarization on spontaneous conversations☆

Dong Wang, Yang Liu

Department of Computer Science, The University of Texas at Dallas, 800 West Campbell Road, Richardson, TX 75080, United States

Received 31 January 2014; received in revised form 2 February 2015; accepted 30 April 2015
Available online 11 May 2015

Abstract

In this study we explore opinion summarization on spontaneous conversations using unsupervised and supervised approaches. We annotate a phone conversation corpus with reference extractive and abstractive summaries for a speaker's opinion on a given topic. We investigate two methods: the first is an unsupervised graph-based method, which incorporates topic and sentiment information, as well as sentence-to-sentence relations extracted based on dialogue structure; the second is a supervised method that casts the summarization problem as a classification problem. Furthermore, we investigate the use of pronoun resolution in this summarization task. We develop various features based on pronoun coreference and incorporate them in the supervised opinion summarization system. Our experimental results show that both the graph-based method and the supervised method outperform the baseline approach, and the pronoun related features can help to generate better summaries.

Keywords: Opinion summarization; Switchboard corpus; Sentiment analysis; Graph-based summarization; Pronoun resolution

1. Introduction

There is a growing interest in sentiment analysis in the Natural Language Processing (NLP) community. Most of the previous research focuses on identifying the polarity or subjectivity of a text document or sentence. However, in many cases people's opinion is mixed or vague, and it is hard to measure using simple scales such as positive, negative or neutral. In the following examples, the first one shows two sentences extracted from product reviews. These short sentences express mixed opinions of pros and cons. The second example is a person's opinion on "gun control", extracted from a conversation transcript. This person is against gun control but agrees that there should be some legislation.

[Example 1] (Ganesan et al., 2010)
iPhone's battery is bulky but it is cheap...
iPhone's battery is bulky but it lasts long
[Example 2]
well, on a scale of one to ten, uh, being ten, no, kind of legislation and zero being, uh, total ban, I probably would lean more towards six or seven.

☆ This paper has been recommended for acceptance by Edward J Briscoe.

E-mail addresses: dongwang@hlt.utdallas.edu (D. Wang), yangl@hlt.utdallas.edu (Y. Liu).

http://dx.doi.org/10.1016/j.csl.2015.04.004
0885-2308/

um, I feel like a total ban on guns is just going to put the guns in the hands the criminals.
I shouldn't say I don't know if it was any worse,
but it certainly didn't get any better.
I think that the law are way too lax.
I, I agree the thing that scares me, uh, though about where I would definitely want some sort of legislation
I don't believe that people should be allowed to carry guns in their vehicles.

For cases shown above, it is hard to determine whether the sentiment is positive or negative. Moreover, users may be more interested in the reasons behind the opinions, rather than a simple answer such as "support or against" or "like it or not". Under these circumstances, an opinion-based summary is able to better represent people's opinion.

Different from generic summarization, an opinion-based summary summarizes people's opinion rather than objective facts. It can be used in summarizing reviews or comments where we care more about what people think about a person or a product. In this study, we investigate opinion summarization on spontaneous conversations. The task 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 is useful for many applications, especially for processing the increasing amount of conversation recordings (e.g., telephone conversations, customer service, round-table discussions, interviews in broadcast programs) where we often need to find a person's opinion or attitude, for example, "what does the speaker think about capital punishment and why?". This kind of questions can be treated as a topic-oriented opinion summarization task.

In this work, we annotate a subset of the Switchboard corpus with human summaries. This is one of the first corpora for this study. We compare two methods that have been widely used in extractive summarization: graph-based and supervised methods. Our system attempts to incorporate more information about topic relevance and sentiment scores. In addition, in the graph-based method we propose to incorporate the dialogue structure information in order to select salient summary sentences. Next we exploit the use of pronoun resolution in spontaneous conversation summarization. We analyze the referent types in spontaneous conversations and study the potential improvement by using pronoun information. Then we propose a variety of pronoun related features and incorporate them in the supervised summarization approach. Our experimental results show that both the unsupervised and supervised methods achieve better results compared to the baseline, and the pronoun features are helpful to extract better summary sentences. In addition, we measure the performance of three automatic coreference resolution systems by comparing with human annotation, and find that all of the three tools cannot perform well on the spontaneous conversation domain for our defined pronoun set. It shows the limit of using automatically generated pronoun resolution results in conversation summarization.

This paper is an extension of our previous work in Wang and Liu (2011). Our new main contributions in this paper are: (i) a new study of sentiment analysis for conversations, including data annotation, automatic classification, and an analysis of relationship between subjective sentences and opinion summaries, (ii) use of a supervised approach and comparison with the graph-based method for opinion summarization, (iii) using both human transcript and speech recognition output for summarization, and (iv) exploration of pronoun information in summarization of conversations.

2. Related work

Research in document summarization has been well established over the past decades. Previous studies have used various domains, including news articles, scientific articles, web documents and 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), and voice mails (Koumpis and Renals, 2005). In general speech domains seem to be more difficult than well-written text for summarization.

In previous work, both unsupervised and supervised methods have been studied for text or speech summarization.

- Unsupervised methods: The Maximal Marginal Relevance (MMR) approach selects the most salient sentences and at the same time avoids redundancy. The salience is measured by the similarity of a sentence and the whole document, and redundancy is measured using the similarity of a sentence and the sentences already selected in the summary. Latent Semantic Analysis (LSA) applies singular value decomposition (SVD) on a term by sentence matrix and then chooses the most important sentences. Graph-based methods (Erkan and Radev, 2004; Garg et al., 2009; Chen

and Metze, 2012) generally represent each sentence as a node and their similarities as edges, then use an iterative method similar to PageRank (Page et al., 1999) to enhance the score of sentences that are more similar to others. Recently there is a growing interest in utilizing the integer linear programming (ILP) framework for summarization. In this method, the sentence selection task in summarization is represented as a global optimization problem that aims to maximize some objective function (e.g., maximizing the selected concepts in (Gillick et al., 2009)) under the summary length constraint. Other global optimization methods have also been investigated, such as using the submodular functions (Lin and Bilmes, 2010).

- Supervised methods: Supervised approaches cast the extraction problem as a binary classification or sequence labeling task. Different classifiers have been adopted, such as support vector machines (SVMs) (Wong et al., 2008; Lerman and McDonald, 2009) and conditional random fields (CRFs) (Shen et al., 2007). Many kinds of features have been also explored, such as sentence length, sentence position, similarity to other sentences, and appearance of topic words. Prior research in speech summarization has also explored using speech specific information, including prosodic features, dialog structure, and speech recognition confidence (Murray and Carenini, 2008; Xie et al., 2009).

In order to provide a summary over opinions, we need to find out which sentences in the conversation contain subjective content. Previous work in sentiment analysis has focused on reviews (Pang and Lee, 2004; Popescu and Etzioni, 2007; Ng et al., 2006) and news resources (Wiebe and Riloff, 2005), with a recent surge on the social media data. Only a handful studies have used conversational speech for opinion recognition, in which some domain-specific features such as prosodic features or conversational structure features were utilized. Murray and Carenini (2009) used a large set of features including n-grams, varying instantiation n-grams, patterns learned from annotated and unannotated data, and conversational features on polarity classification and subjectivity detection. They found that their proposed features are able to improve the baseline using only trigram features. Raaijmakers et al. (2008) compared the features from two modalities: words and acoustics. Their results showed that character n-grams outperform other features and prosodic features are least valuable, while using all the features yields the best result.

Opinion summarization has been studied in some previous research. For example, Hu and Liu (2004) and Nishikawa et al. (2010) explored opinion summarization in customer reviews. Paul et al. (2010) summarized contrastive viewpoints on multiple types of opinionated text, such as surveys and editorials. Meng et al. (2012) summarized opinions in Twitter messages on specific topics, by exploiting the #hashtags, which is a specific marker used to indicate topics in Twitter. Opinion summarization was also 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. Opinion question answering can also be cast as an opinion summarization task. Stoyanov et al. (2005) applied a subjectivity filter based on traditional Question Answering (QA) systems to generate opinionated answers. Balahur et al. (2010) answered 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. However, opinion summarization in spontaneous conversation is seldom studied.

To utilize coreference information in summarization, Azzam et al. (1999) selected the "best" coreference chain using a variety of criteria to represent the main topic of a text. Bergler et al. (2003) used the length of the noun phrase coreference chain to select the most important entities to form a 10-word summary. Steinberger et al. (2005) showed that an LSA-based method using anaphora resolution helps summarization, but simply substituting with pronouns hurts the performance. Stoyanov and Cardie (2006) used coreference resolution to group the opinions from each source for opinion summarization. All of these previous studies used written text. There has not been much work investigating the use of pronoun information for summarization in speech domains, especially conversational speech.

There are several differences between our work and the previous ones. First, we investigate opinion summarization in spontaneous conversations, a domain that has not been explored much for sentiment analysis or summarization. Second, we combine conversational structure into the graph-based summarization method to better represent the conversation. Third, we exploit pronoun information for summarization by incorporating pronoun related features in the supervised approach.

3. Summary corpus creation

Though there are many annotated data sets for the research of speech summarization or sentiment analysis, there is no corpus available for opinion summarization on spontaneous speech. Therefore for this study, we create a new pilot

Table 1
Corpus statistics: topic description, number of conversations in each topic, average length (number of sentences) of all the conversations, and standard deviation (SD).

Topic	#Conv.	Avg length	SD
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		

data set using a subset of the Switchboard corpus (Godfrey and Holliman, 1997). 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 corpus, we selected 88 conversations from 6 topics for this study. Table 1 lists the number of conversations in each topic, the average number of sentences of all the conversations,[1] and standard deviation of the conversation length.

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 (from one of the three annotators).

The annotators have access to both transcripts and the 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 sentences (no minimum limit) in the transcripts for each speaker, from which their abstractive summary is derived. The selected sentences 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 average compression ratio of the extractive summaries and abstractive summaries as well as their standard deviation. Because the sentence length varies a lot in conversations, we use words instead of sentences as the units when calculating the compression ratio.

[1] We use "sentences" in this work. These are obtained based on the dialog act annotation in the corpus. Note that sentences in conversational speech are different from those in the written text.

Table 2
Compression ratio and standard deviation (SD) for extractive and abstractive summaries.

	Avg ratio (%)	SD
Extractive summaries	0.26	0.13
Abstractive summaries	0.13	0.06

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

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	$\alpha = 0.79$	

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) 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, which is more suitable for measuring human agreement for interval data than Cohen's kappa.[2]

We notice that the inter-annotator agreement for extractive summaries is comparable to other speech summary annotation (Liu and Liu, 2008). The agreement on abstractive summaries is much lower than extractive summaries, which is as expected. The agreement for the overall opinion annotation is similar to other opinion/emotion studies (Wilson, 2008), but slightly lower than the level recommended by Krippendorff for reliable data ($\alpha = 0.8$) (Hayes and Krippendorff, 2007). This shows that 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. This also suggests 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. Sentiment categorization for opinion summarization

4.1. Sentiment annotation

Since our task is opinion summarization, intuitively we need to determine the sentiment for each sentence. In order to study the relation between opinionated sentences and opinion summaries, we first annotate the sentiment on the sentence level in a subset of the corpus above. We use the 18 conversations that have multiple summary annotations. For each sentence, we label it as one of the following categories:

- subjective question: questions that are used to elicit the opinion of the other speaker.
  [**Example:**]
  - What do you think?
  - How do you feel?
  - Should we go ahead and explore?
- subjective statement: sentences that contain opinion of the speaker himself or from others. The opinion can be positive, negative, or mixed. The sentence may express agreement or disagreement.

[2] To calculate Krippendorff's α, we used the following difference function for the 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.

Table 4
Inter-annotator agreement on sentiment annotation.

Category	Kappa
Overall	0.472
Subjective question	0.687
Subjective statement	0.459
Uncertainty	0.661
Objective-polar	0.118
Objective	0.517

Table 5
Number of sentences in each category.

Category	# Sentences
Subjective question	60
Subjective statement	1393
Uncertainty	66
Objective-polar	131
Objective	1714

[Example:]
- I do not understand why we do not enforce the law we have
- That would certainly help
- I am sure

- uncertainty: the speaker is uncertain about something.
  [Example:]
  - oh, I do not know
  - I am not really sure what texas law, I think there be a check for felony, on your record
- objective-polar: sentences that describe positive or negative factual information about something.
  [Example:]
  - I fortunately, I have never be in that circumstance
  - There is one guy that we test for preemployment for drug who absolutely swear up and down that it shows that he is using amphetamine
- objective: objective sentences with no polarity contained. It can be a statement or a question.
  [Example:]
  - I have opportunity in high school to work in some program.
  - Where do you live?
  - Do you have any children?

In order to study the inter-annotator agreement, we have 4 conversations from different topics annotated by 3 annotators. The number of sentences in the 4 conversations is 1039. Table 4 shows the inter-annotator agreement, measured using Fleiss' kappa (Joseph, 1971). The agreement on category "subjective question" and "uncertainty" is quite high because they are easy to identify, but is very low on category "objective-polar", as it is always confused with "subjective statement" or "objective" class. For these 4 conversations, we use the majority vote to select the best label and use it in the following analysis and experiments. For the other conversations we have only one annotation.

Table 5 shows the number of sentences in each category based on the annotated conversations. We can see that more than half of the sentences are objective, and the next largest category is subjective class. These two categories take over 90% of all the sentences.

Table 6 shows the percentage of each category in the extractive reference summaries. Because these 18 conversations have multiple extractive reference summaries (from 3 annotators), we use the majority vote to select summary sentences in this analysis. 319 sentences are selected as gold summary sentences. Different from the class distribution in the

Table 6
The percentage of each sentiment category in reference summaries.

Category	Percentage in summaries (%)
Subjective question	0.3
Subjective statement	79.3
Uncertainty	2.5
Objective-polar	5.6
Objective	12.2

entire dataset (Table 5), 79.3% of sentences in reference summaries are subjective, followed by the objective class that takes 12.2%. This shows the importance of detecting subjective content for opinion summarization.

4.2. Sentiment categorization on switchboard corpus

Given the evidence that the subjective sentences take about 80% in the reference summary sentences, detecting these sentences is needed for summarization. We cast the subjectivity detection problem as a binary classification problem with two classes: "subjective" and "non-subjective". "Subjective" class include sentences labeled as "subjective statement" in the 5 sentiment categories we defined, and "non-subjective" class include sentences labeled as one of the rest 4 categories ("subjective question", "uncertainty", "objective-polar", and "objective"). We choose to use the maximum entropy classifier in this study, since it has achieved competitive performance in many classification tasks (Berger et al., 1996; Nigam, 1999), and it can provide a confidence score that we will use in selecting summary sentences.

For the features in the supervised subjectivity classifier, in addition to the n-gram lexical features which have been widely used in sentiment analysis and are very strong baseline features, we explore a few domain specific features. The following lists all the features we use.

- Lexical features: n-gram features (unigrams to trigrams).
- The number of words in the current sentence and the relative position of the current sentence in the whole conversation, as they always indicate the importance of the sentence. These are represented as binary features.
- Number of sentences in the current turn. More sentences in a turn may indicate more important content. For example, in the following conversation there are 4 sentences from Speaker A in the same turn.
  - *B: That's right.*
  - *A: I don't know.*
  - *A: Coming from Texas you're probably, -*
  - *A: I shouldn't make stereo types,*
  - *A: but gun control is probably frowned on quite a bit down there I would think.*
  - *B: Yeah.*
- Number of words in the previous sentence. If the current sentence is the first one in the dialogue, we use its own length.
- A binary feature to indicate if the current sentence is the first one in the dialogue, because the first sentence in a dialogue is always greeting or self-introduction, and is not related to the content of the conversation.

Since we have limited data to train the subjectivity classifier for the conversation domain, we decided to use some domain adaptation approaches in order to leverage labeled data from other domains. The Switchboard conversation data is our target domain in this study. The source domain data we considered includes the movie data (Pang and Lee, 2004) and news articles in the MPQA corpus (Wilson and Wiebe, 2003). We evaluated two domain adaptation approaches:

- Adapt1: We train a model on the source domains and use its prediction as an additional feature when building the model on the target domain.

Table 7
Sentiment categorization results: F scores for "subjective" and "non-subjective" class, and the overall accuracy.

Feature set	Sub F (%)	Non-sub F (%)	Accuracy (%)
Lexical features	54.64	71.55	65.03
All features w/o domain adaptation	56.62	72.22	66.13
All features + adapt1	**58.64**	**72.75**	**67.14**
Lexical features + adapt2	56.08	70.66	64.82

- Adapt2: We use the domain adaptation method proposed by Daume (2007). This domain adaptation method expands the feature set of both the source and the target domain to an augmented feature set including the domain specific features and the general features. The augmented features for the source (ϕ^s) and target domains (ϕ^t) are defined as:

$$\phi^s(x) = < x, x, 0>$$
$$\phi^t(x) = < x, 0, x>$$

  where the three vectors in the expanded feature space represent the common, source-specific and target-specific components. x is the vector in the original feature space, and 0 is the zero vector. The model is built on a combined dataset including the source data and target domain data. Note that this method assumes that the original feature set from the source domain and the target domain should be the same. Since we do not have feature "position" in the two source domains, and the sentence length range in written text and speech is quite different, we only use the n-gram features in this domain adaptation approach.

We perform a 18-fold cross validation by using one conversation as the test set at each run. Table 7 shows the F-measure on "subjective" and "non-subjective" classes and the overall accuracy using different feature sets and adaptation methods. Using all the features with "adapt1" domain adaptation method yields the best results (comparing to "lexical features", $p = 0.017$ using paired t-test). The other domain adaption method (adapt2) does not help much on this problem, possibly because it only uses n-gram lexical features and cannot use any domain specific features. Thus in the following summarization process we use all the features with "adapt1" domain adaptation for subjectivity detection.

5. Opinion summarization methods

5.1. Graph-based approach

Graph-based methods have been widely used in document summarization. In this approach, a document is modeled as a graph, where each node represents a sentence. 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 (e.g., (Erkan and Radev, 2004)) showed that this method can effectively extract important sentences from documents. The framework we use in this study is similar to the query-based graph summarization system in (Zhao et al., 2009). In their task, they only consider the similarity between each pair of sentences and the relevance between query and each sentence. In our work, we also consider sentiment and topic relevance information, and propose to incorporate information obtained from dialog structure. The score for a sentence s, which represents its importance in the conversation, is based on four factors: its content similarity with all the other sentences in the dialogue, the connection with other sentences based on the dialogue structure, the topic relevance, and its subjectivity, that is:

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

where C is the set of all the sentences in the dialogue.

- $sim(s, v)$ is the cosine similarity between two sentences s and v.
- $REL(s, topic)$ measures the topic relevance of sentence s. It is the sum of the topic relevance of all the words in the sentence. We only consider the content words for this measure. The content words are identified using the TreeTagger toolkit.[3] To measure the relevance of a word to a topic, we use Pairwise Mutual Information (PMI):

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

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 sentences in the same dialog, and the topic is the same for all the sentences, 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:

$$REL_{orig}(s, topic) = \sum_{w \in s} PMI^+(w, topic)$$

$$REL(s, topic) = \frac{REL_{orig}(s, topic) - Min}{Max - Min}$$

- $sentiment(s)$ indicates the probability that sentence s contains opinion. To obtain this, we trained a binary maximum entropy classifier on the training set that includes all the conversations annotated with sentiment labels, except the conversation that is being processed. We use the best feature set described in Section 4. We use each sentence's probability of being "subjective" predicted by the classifier as its sentiment score.

In addition to the three scores above, we introduce new connections $ADJ(s, v)$ to model the dialog structure. It is a directed edge from s to v, defined as follows:

(A) 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 $1/dis(s, v)$, i.e., $ADJ(s, v) = ADJ(v, s) = 1/dis(s, v)$, where $dis(s, v)$ is the distance between s and v, measured by the number of sentences they are separated away from each other. This way the sentences in the same turn can reinforce each other – if one sentence is important, then the other sentences in the same turn are also important.

(B) If s and v are from the same speaker, and separated only by one sentence 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, i.e., $ADJ(s, v) = ADJ(v, s) = 1$.

(C) 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, i.e., $ADJ(s, v) = 1$. 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 that, if the score of a question sentence is high, then the answer's score is also boosted.

(D) If s and v form an agreement or disagreement pair, then there is an edge from v to s with weight 1, i.e., $ADJ(v, s) = 1$. The agreement/disagreement pairs are also determined by simple rules: sentence v contains the word "agree" or "disagree", and s is the previous sentence from a different speaker. The reason for adding this connection is similar to the above question-answer pairs. Note that the rules used to detect the agreement/disagreement pair (and the question-answer pair above) are quite simple. It is likely that many such pairs are missed using our keyword-based method. We plan to investigate using better adjacency pair detection approaches in our future work.

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

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

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 standard graph-based methods, summarization is conducted for each speaker separately. Sentences 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 sentences.

5.2. Supervised approach

The supervised method casts the summarization problem as a classification or regression problem. We train a model using the training set, where each sentence has a label to indicate whether it is a summary sentence or not. This model is used to predict the score for each sentence in the test conversations. Then we select top-ranked sentences according to the predicted scores as summaries. In this work we use four features for opinion summarization:

- *REL*(*s*, *topic*) is the same as the one used in the graph-based approach.
- *sentiment*(*s*) is the same as the one used in the graph-based approach.
- *sim*(*s*, *D*) is the cosine similarity between sentence *s* and all the sentences in the dialogue from the same speaker, *D*. It measures the relevance of *s* to the entire dialogue from the target speaker. This score is used to represent the salience of the sentence. 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.
- *length*(*s*) is the sentence length. This score can effectively penalize the short sentences that 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].

We also propose two variations in order to improve the summarization system performance.

(I) **Sampling**

When treating the extractive summarization as a binary classification task, there is an imbalanced data problem – in our data set, only 319 sentences are selected in the reference summary from the total of 3366 sentences. In this case, the machine learning models tend to produce high accuracy over the majority class, but poor accuracy over the minority class (Maloof, 2003). In this work we apply an up-sampling method to address the imbalanced data problem – we replicate the summary sentences by n times ($n = 5$ in this work) during training. This is expected to increase the recall of the positive class (summary sentences), though it may also overfit these positive examples.

(II) **Score normalization**

A common way to select summary sentences on the test set is to select the top ranked sentences based on the scores (which represent the probability that the sentence is in summary) from the classifier until the length limit is reached. In this work, rather than just using the original scores from the classifier, we propose to normalize them by the sentence length, shown below. We empirically set w to 30. This is expected to address some preference bias towards the long sentences and select more informative sentences under the same length constraint.

$$new\_score = \frac{score}{\log(sent\_length + w)} \quad (w = 30)$$

6. Summarization results using graph-based and supervised approaches

6.1. Experimental setup

We use the 18 conversations annotated by all the three annotators as the test set. The rest 70 conversations are used as development set to tune the parameters (determining the best combination weights) in the graph-based method, or

as the training set in the supervised method. In preprocessing we applied word stemming on all the words. We perform extractive summarization using different word compression ratios (ranging from 10% to 25%). We use human annotated dialogue acts (DA) as the sentences for extractive summarization. We evaluate our systems on both manual transcripts and automatic speech recognition (ASR) output. The word error rate of ASR output on this corpus is 31.39%. We align the ASR output with human transcripts based on edit distance to get the dialogue act units in ASR output. The system-generated summaries are compared to human annotated extractive and abstractive summaries. We use ROUGE as the evaluation metrics for summarization performance. In calculating ROUGE F-score, we put more importance on recall, using $\alpha = 0.8$ in Eq. (3), which is the same as (Lin et al., 2012).

$$F = \frac{Precision \times Recall}{\alpha \times Precision + (1 - \alpha) \times Recall} \tag{3}$$

We compare our methods to two systems. The first one is a baseline system, where we select the longest sentences 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, and compute their average. This can be considered as the upper bound of our system performance.

6.2. Results of graph-based and supervised approaches

In the graph-based approach, we used the grid search method to obtain the best combination weights on the development set. The best parameters we found are $\lambda_{sim} = 0$, $\lambda_{adj} = 0.4$, $\lambda_{rel} = 0.2$, $\lambda_{sent} = 0.4$. It shows that the similarity between each sentence pair is not useful. This is different from graph-based summarization systems for text domains. A similar finding has also been shown in (Garg et al., 2009), where using similarity between sentences does not perform well in conversation summarization. One possible reason is that, in Switchboard conversations, what people talk about is diverse and in many cases there are not many common words between two sentences.

In the supervised approach, we apply the SVM regression model for the extractive summarization task. For training, the sentences that are selected by annotators are assigned 1 as the target value; otherwise it is 0.

Fig. 1 shows the ROUGE-1 F-scores comparing to human extractive and abstractive summaries using different compression ratios for different systems. The results are shown for the human transcripts and ASR test conditions for the automatic summarization systems. Similar patterns are observed for other ROUGE scores such as ROUGE-2 or ROUGE-L, therefore they are not shown here.

On manual transcripts, we can see that our two summarization methods (graph-based and supervised methods) clearly outperform the max-length baseline system. When comparing to the extractive reference summaries (Fig. 1(a)), the graph-based approach works better than supervised approach for lower compression ratios, and the supervised method works slightly better for higher compression ratios. When comparing to the abstractive reference summaries (Fig. 1(b)), the graph-based method works better than the supervised approach. Both methods outperform the baseline system for different compression ratios (significance level of $p < 0.02$ in paired t-test).

Fig. 1 also shows that when using ASR output, overall there is a performance degradation compared to using human transcripts, mainly because of ASR errors. The graph-based approach performs better than the baseline (max-length_asr) when comparing to both extractive and abstractive reference summaries. That demonstrates the robustness of the graph-based approach. However, the supervised approach has similar performance compared to the baseline approach for some conditions.

We also compared the performance of the supervised approach to the same approach without using sampling or score normalization, and found that using them improved the ROUGE score when compared to abstractive reference summaries but did not have much difference when compared to extractive reference summaries.

6.3. Analysis of dialogue structure in graph-based approach

Since we have seen that the graph-based method achieves the best performance, we conduct a further analysis with a focus on the effect of the dialogue structure we introduce in the graph-based summarization method. To study this, we compare two configurations: $\lambda_{adj} = 0$ (only using REL score and sentiment score in ranking) and $\lambda_{adj} = 0.4$. We generate summaries using these two setups on manual transcripts and compare with human selected sentences. Table 8 shows

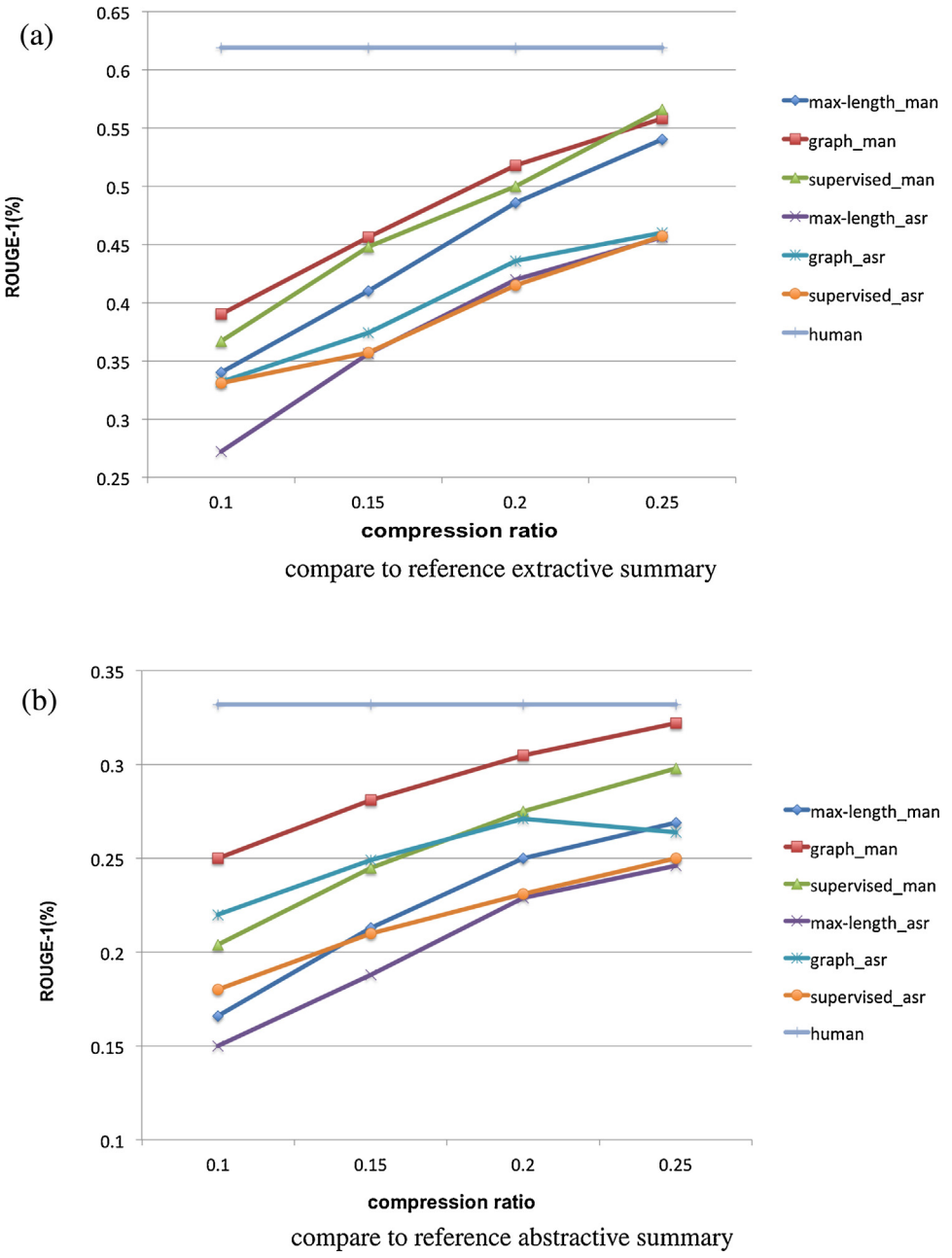

Fig. 1. ROUGE-1 *F*-scores compared to extractive and abstractive reference summaries for different systems: max-length, graph-based method, supervised method, and human performance, using manual transcripts and ASR output.

the number of false positive instances (selected by the 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 a sentence as positive if it is selected by at least one annotator. This results in a large number of reference summary sentences (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 sentences 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 0.2.

Table 8
The number of false positive (FP) and false negative (FN) instances using the graph-based method with $\lambda_{adj} = 0$ and $\lambda_{adj} = 0.4$ for different compression ratios.

Ratio	$\lambda_{adj} = 0$		$\lambda_{adj} = 0.4$	
	FP	FN	FP	FN
0.1	41	589	37	583
0.15	70	559	63	549
0.2	104	529	103	529
0.25	141	497	137	493

The following shows an example, where the third sentence is selected by the system with $\lambda_{adj} = 0.4$, but not by $\lambda_{adj} = 0$. This is partly because the weight of the second sentence is enhanced by the question-answer pair (the first and the second sentence), and thus subsequently boosting the score of the third sentence.

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 : Itwouldprobablybesomewherecloserto, uh, lesscontrolbecauseIdon′tsee, –

7. Exploit pronoun resolution for opinion summarization in conversations

7.1. Pronoun annotation

To study the relation between pronoun resolution and summarization, we annotate the pronouns in the 18 Switchboard conversations that have summary annotation from three different annotators.

Since our goal is to evaluate the impact of coreference resolution on speech summarization, we chose to focus on 11 pronouns that we think might be helpful for the summarization task. These are: [**it, they, them, their, he, his, him, she, her, this, that**]. We do not consider other pronouns, such as the first/second personal pronouns (I, me, you), reflective pronouns (himself, themselves), negative pronouns (nobody). We use Amazon Mechanical Turk to collect the annotations. For each conversation, we give the annotator the whole conversation transcripts, where all the pronouns are highlighted. For each pronoun, the annotator is asked to do three things:

- select the referent category for that pronoun from the following list: noun phrase (NP), verb phrase (VP), sentence, referential-other, not referential.
- if "NP", "VP" or "Sentence" is selected, write down the indices of the sentence where the referred entity is extracted. If the referred entity appears repeatedly in more than one sentence, use the one closest to the current line; and if the referred entity is a sentence/long phrase which crosses multiple lines, select all of them.
- write down the referred entity using the original words from the previous sentence.

The following explains the pronoun categories in more details and shows some examples (pronouns and their referred entities shown in bold).

(A) NP: Noun Phrase.
  - *B: I'm very much in favor of* **gun control**.
  - *A: I'm against* **it**
(B) VP: Verb Phrase.
  - *A: I hate* **having to be in charge of someone else's life**.
  - *B: Sure*, **it***'s a big responsibility*.
(C) Sentence: The pronoun refers to a sentence that contains a subject, a verb with/without an object. For example,
  - *A: well*, **they're not rigid enough here, in Texas**.
  - *B: Yeah*, **that***'s true.*
  - *A: But don't tell the N R A I said* **that**.

Table 9
Inter-annotator agreement on pronoun coreference annotation.

Category	Kappa
Overall	0.361
NP	0.455
VP	0.173
Sentence	0.215
Not Referential	0.489
Referential-other	0.167

Table 10
Percentage (%) of five categories of pronouns.

NP	38.01
VP	3.34
Sentence	5.84
Referential-other	21.95
Not referential	30.87

(D) Referential-other: If the pronoun refers to some previous context, but it is hard to find a clear and explicit antecedent, or the pronoun refers to a word/phrase that is implied, but not stated, such as "they" in the following example:
- *Well, you know, uh, now here's something that, uh, first occurred to me when* **they** *started having all these problems with these automatic, uh, weapons.*

(E) Not referential: The pronoun is an expletive or dummy pronoun, such as "it" in the following two examples:
- **it** *rains outside.*
- **it** *is obvious that...*

Each conversation is annotated by 5–9 annotators, and each takes about 30 minutes on average to accomplish. All the 18 conversations are annotated within a week. We rejected several submissions with low qualities (which have many missing fields or many obvious wrong answers) and kept only 5 annotations for each conversation.

We evaluate the inter-annotator agreement with Fleiss' kappa (Joseph, 1971) using the 5 annotations from each conversation. The overall kappa value and kappa on each category are shown in Table 9. The annotators mostly disagree on "VP" and "sentence". The agreement for the "NP" and "Not Referential" is much higher. To measure the agreement on the referred entity, we calculate the cosine similarity between each pair of annotations for the same pronoun, then take the average. This results in an agreement of 0.55. In many cases the annotations are different simply because they use different phrase boundaries, for example, "the guy" vs. "the guy who robs seven eleven". To select the best annotation as gold standard, we measure the pairwise cosine-similarity between each pair of referred entities, and select the one which has the highest similarity with the rest of the annotations, and use its reference category as well.

In the 18 conversations, there are 2056 pronouns annotated in total. Table 10 shows the distribution of the 5 categories. Obviously "Referential-other" or "Not referential" pronouns cannot benefit the summarization task much since even humans cannot tell what they refer to. For the rest of the pronouns, 13% of them refer to the phrase/sentence in the same line, which we expect are not helpful in summarization since the information that the antecedents represent is already contained in the same sentence.

Without considering the above types, 819 pronouns out of 2,056 remain. We are going to focus on these to evaluate their impact on summarization. Therefore in the following of the paper, "pronoun" only denotes the pronouns we use in this study, that is, pronouns referring to "NP", "VP" and "Sentence" in a different line.

7.2. Relation between summarization and pronoun resolution

We investigate whether and to what extent pronoun resolution helps summarization under the framework of topic-oriented opinion summarization. First we perform some analysis to show whether and how often reference summary sentences contain pronouns, and whether automatic systems fail to identify such summary sentences.

Table 11
The percentage of sentences without pronouns, sentences with pronouns and selected by the automatic summarization system, and sentences with pronouns but not selected by the automatic system.

Sentence with pronoun but not selected by system	22.26%
Sentence with pronoun and selected by system	11.60%
Sentence without pronoun	66.14%

Each of the 18 conversations in the corpus has annotated abstractive and extractive summaries for each speaker about his/her opinion or attitude on the given topic. Each conversation was annotated by 3 annotators. For the following analysis, we use their majority vote to create the gold standard extractive summaries, i.e., if a sentence is selected by two or more annotators, it is included in the reference summary. For the automatic summaries, we generate the system summaries using the supervised approach described in Section 5.2 with a compression ratio of 0.25.

Table 11 shows the breakdown of the sentences for the reference and automatic summaries with respect to pronouns. In the reference summaries, 108 out of 319 sentences contain pronouns, and most of them are not selected by the system. A further examination shows that over 90% of the sentences with pronouns in reference summaries contain "NP" references. This analysis shows the potential improvement on summarization by integrating pronoun resolution, especially the NP coreference resolution.

The following shows several false negative examples (sentences are in the reference summaries, but not in the system generated ones). In the examples, we also include their annotations. The tag "value" shows the referred phrase, "category" is the reference type, and "refline" shows the line where the referred phrase is from.

- *Um, personally, I don't have a problem with <pron value="drug testing" category="NP" refline="2">**it**</pron>.*
- *because I'm against <pron value="gun control" category="NP" refline="1 3">**it**</pron>.*
- *we need to put restrictions on <pron value="more complex weapons" category="NP" refline="14">**them**</pron>.*

7.3. Employ pronoun coreference resolution in opinion-oriented topic summarization

Since NP reference is the most frequent one in the summary sentences (compared to VP and sentence categories), and NP reference is also the focus of most of current coreference resolution systems, we only exploit NP references in our summarization system. In this work, we use the supervised summarization approach since it is more flexible in incorporating pronoun information (as additional features).

We use the four scores used in the supervised approach (Section 5.2) as the baseline features. In order to use the information contained in the NP that the pronoun refers to, we develop additional pronoun related features, listed below:

- Reference context features: If a pronoun in the current sentence refers to an entity in another sentence r with a topic score of $REL(r, topic)$ and sentiment score of $sentiment(r)$, then we add the information from that sentence weighted by the topic score of the NP:

$$\begin{aligned} context\_fea\_t &= REL(r, topic) \times TOPIC(NP) \\ context\_fea\_s &= sentiment(r) \times TOPIC(NP) \end{aligned} \tag{4}$$

  where $TOPIC(NP)$ is the topic score of the referred NP, which is calculated as the average relevance score of all words contained in this NP (Eq. (5)). If the current sentence refers to multiple NPs, we use the maximum score for the feature.

$$TOPIC(NP) = \frac{\sum_{w \in NP} PMI^{+}(w, topic)}{|NP|} \tag{5}$$

- Reference feature: if a sentence contains an NP that other pronouns refer to, then we expect that the more times the NP is referred to, the more important the sentence containing the NP is. We add a feature to represent the importance of such NPs.

$$ref\_fea = \sum_{NP_i \in sentence} ref\_freq(NP_i) \times TOPIC(NP_i)$$

where $ref\_freq(NP_i)$ is the number of times that NP_i in the current sentence is referred to.

- Coreference chain feature: we extract the coreference chain for all the referred NP, and use its length weighted by the topic relevance of the referred NP as a feature for the current sentence.

$$chain\_fea = chain\_length \times TOPIC(NP)$$

where *chain_length* is the length of the coreference chain where the current sentence is involved, and $TOPIC(NP)$ is the topic score of the NP. If there exist multiple coreference chains, we use the maximum score for this feature.

After the sentence selection process on the test set, when generating the summaries, we try to deal with the dangling pronouns. If a summary sentence contains a pronoun referring to an NP phrase in another sentence that is not selected in the summary, and the NP does not appear in the previous sentence in the summary, then we use the replaced corpus (i.e., replace the pronoun with the corresponding NP). Otherwise, we use the original sentence.

7.4. Experiments

7.4.1. Experimental setup

We apply the SVM regression model for the supervised summarization system. Because we have only pronoun annotation on 18 conversations, we perform 18-fold cross validation by using 17 conversations for training at each run. For training, we generate the target value according to the number of votes from the annotators: if the sentence was selected by all the three annotators, the value is 1; if it was selected by two of them, the value is 0.8; if only one of them selected it, the value is 0.5, otherwise 0. Our preliminary experiments showed that using these weighted target values performed better than using binary target values (e.g., based on the majority vote annotation).

In evaluation, we measure the system performances using ROUGE F-score averaged on 36 summaries (one summary for each of the two speakers in a conversation). In addition, we calculate the F-score of sentence labeling accuracy by comparing the system selected summary sentences with gold standard summaries (derived by the majority vote). Again, we use $\alpha = 0.8$ (Eq. (3)) for the F-scores. We show the averaged score over all the 36 summaries in the results.

We compare our method with the following two systems:

- Baseline: This is the same approach as used in Section 5.2, where we use the baseline features (similarity score, topic score, sentiment score, and length score calculated using the original corpus), without any pronoun information, and use up-sampling and score normalization.
- Baseline_replace: This is similar to Baseline, but we add the pronoun replacement process when necessary in generating the final system summaries (described in Section 7.3). We use this in order to test if resolving the pronouns in post-processing (after selecting summary sentences) is sufficient for summarization.

Our system (SYS_PRON) is as described in Section 7.3: we use the baseline features and pronoun related features, together with the pronoun replacement to deal with dangling pronouns after the summary sentences are selected. None of the above two compared supervised systems uses pronoun related features in the summarization model, which allows us to study the effect of the pronoun features. We use pronoun resolution annotation collected from AMT in our summarization system with the goal of showing the best performance we can expect from incorporating pronoun information. To avoid the compound factors of ASR errors, we evaluate summarization performance on the human transcripts condition.

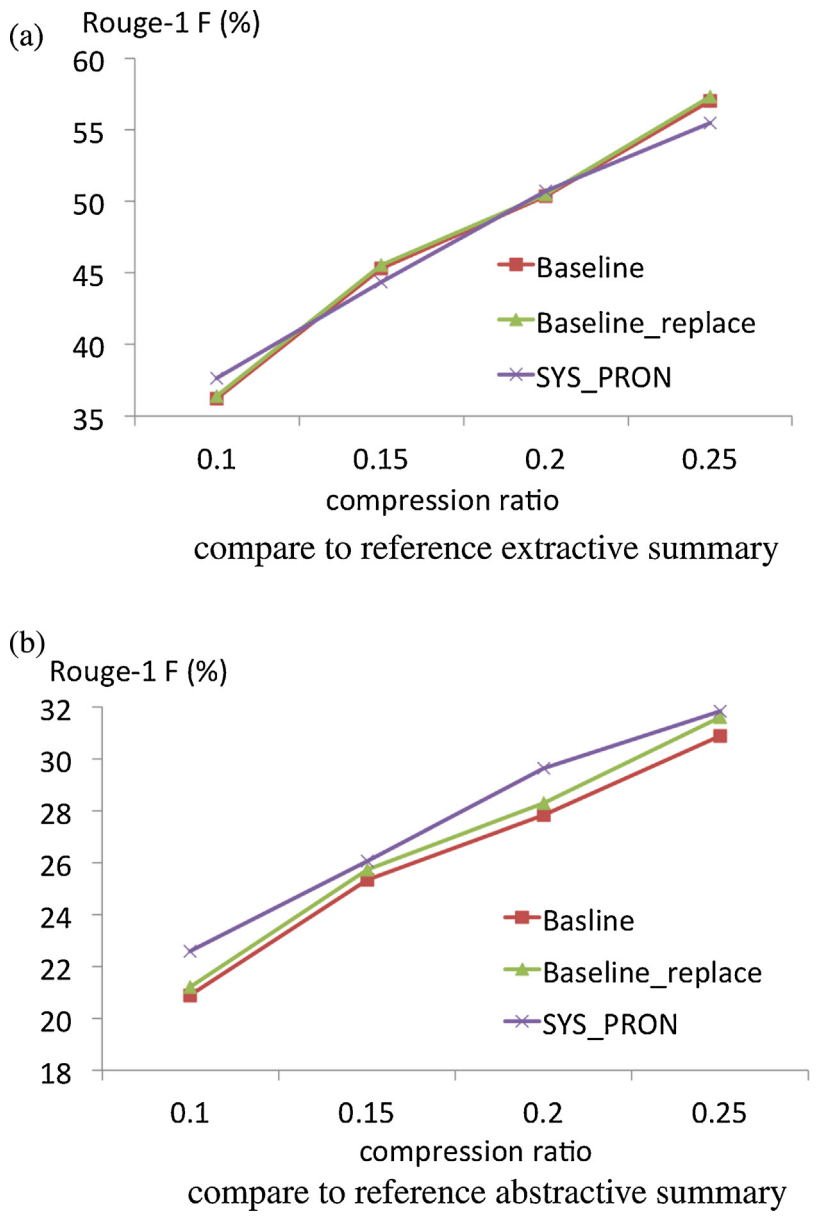

Fig. 2. ROUGE-1 *F*-scores compared to extractive and abstractive reference summaries for different systems: baseline and our system using pronoun information.

7.4.2. Summarization results using pronoun coreference annotation

Fig. 2 shows the ROUGE-1 *F*-score when comparing to extractive and abstractive reference summaries for different systems using different compression ratios. When compared to extractive reference summaries, there is no significant difference among the three systems. SYS_PRON performs better than others on the lowest compression ratio 0.1, but has the worst performance on the highest ratio 0.25. Because the extractive reference summaries are extracted from the original conversation with pronouns, it is not quite appropriate to compare our system output that replaces the pronouns with their referents. Thus we focus on the comparison using the abstractive summaries as references (Fig. 2(b)). We can see the system Baseline_replace yields a gain over Baseline, which shows that using pronoun resolution for post-processing is helpful. SYS_PRON achieves the best result across all the compression ratios from 0.1 to 0.25, indicating the benefit of using the pronoun features in the supervised summarization model. Using a paired t-test, we found that

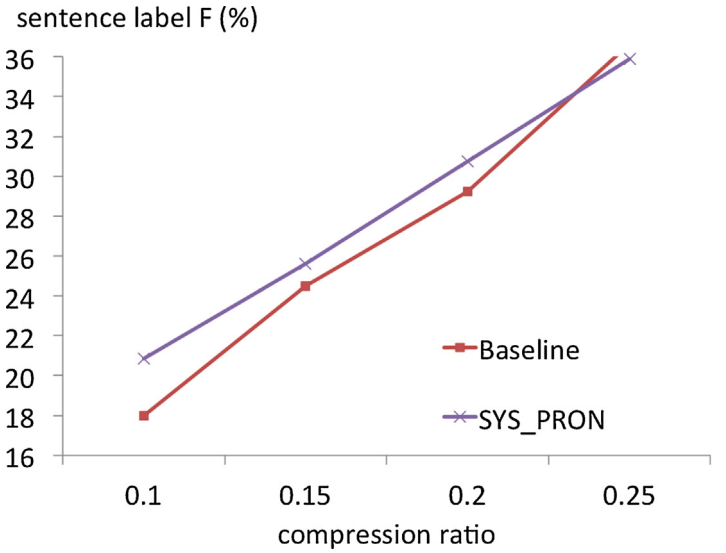

Fig. 3. Sentence label *F*-score (%) using the baseline system vs. the one with pronoun features.

the significance level is $p = 0.016$ when comparing SYS_PRON vs. the baseline system, and $p = 0.06$ for SYS_PRON vs. baseline_replace.

Fig. 3 shows the sentence label *F*-score. Note that Baseline_replace is omitted because it selects the same sentences as Baseline. We can see that SYS_PRON achieves the highest sentence labeling *F*-score for most conditions, except for the highest compression ratio of 0.25, where it is slightly lower than Baseline.

It is interesting that the three systems do not show as much difference on the ROUGE-1 score when comparing with the extractive reference summaries as what is observed when using the sentence labeling accuracy. We also compared the two scores on each summary and found that the reason for this is because the sentence level comparison is harsh. In an extreme case, if no correct summary sentence is selected by the system, the sentence level *F*-score is 0, but it can still get a reasonable ROUGE *F*-score based on word/n-gram units.

The above results demonstrate that pronoun coreference resolution not only can serve as post-processing, but also helps to select better summary sentences, especially when the compression ratio is low.

We also conducted an experiment where we use only the baseline features (similarity score, topic score, sentiment score, and length score) on the conversations with pronoun replacement (i.e., pronoun replacement as a pre-processing step), but it has similar results as the baseline approach, which shows that the pronoun features are critical in improving summarization performance. The following shows some summary sentences that are not selected by Baseline_replace but selected by SYS_PRON when using a compression ratio of 0.2 (we also show the pronoun annotation information in the example). Again, the difference between the two systems is the use of the pronoun related features in SYS_PRON.

- *Because <value="the current generation of space craft" category="NP" refline="7">**they**</pron> have not been proved as reliable as <pron value="the current generation of space craft" category="NP" refline="7">**they**</pron> should have.*
- *I am very much in favor of **gun control**.*

The first sentence is from a conversation about "whether to support space program". It is selected in our system because it contains the pronoun that refers to an NP with a high topic score. The second sentence is from a conversation about "whether to support gun control". It is correctly selected because "gun control" is referred many times throughout this conversation.

Finally we conduct a human evaluation to see if our system is able to improve the human readability. Five conversations were randomly selected for this. Each conversation has 2 summaries, corresponding to the two speakers. Two evaluators judged the summaries generated from the three systems: baseline, baseline_replace and SYS_PRON, using a score ranging from 1 to 5 (higher indicates better readability). Table 12 shows the average score of each system. We

Table 12
Human evaluation of baseline, baseline_replace, and SYS_PRON.

System	Human evaluation score
Baseline	2.0
Baseline_replace	2.4
SYS_PRON	2.7

Table 13
Precision, recall and *F*-score (%) of coreference resolution systems on pronouns.

System	Precision	Recall	*F*-score
Cherrypicker	67.03	6.17	11.31
Reconcile	45.50	9.21	15.32
Stanford	35.27	8.60	13.83

can see that overall these scores are rather low. This is consistent with previous studies, showing extractive summaries from conversations are not coherent and hard to understand. However, among the three systems, our system is the most favorable one. The significance level is $p = 0.008$ when comparing our method to the baseline, and $p = 0.019$ when comparing to baseline_replace.

7.4.3. Using automatic coreference resolution tools

Given the promising result of using pronoun resolution for the summarization task, a natural question is if we can apply the current pronoun resolution tools for this purpose. We thus evaluate the feasibility of using three state-of-the-art automatic coreference resolution tools: "Cherrypicker" (Rahman and Ng, 2009), "Reconcile" (Stoyanov et al., 2010), and "Stanford coreference resolution tool" (Lee et al., 2011). All of them have reported good performance on the ACE corpus, which contains both written text and speech transcripts.

Because of different format and scope of their system output and our annotation, we first transform their system results to a unified format. First we limit the scope within the 11 pronouns we annotated, and then make each pronoun to refer to the nearest previous non-pronoun phrase if they appear in the same coreference chain. In their results, there is no reference category ("NP", "VP", etc.), so we treat all of them as "NP" reference. Because of the difference of task definition and our purpose of using pronouns, we only care about whether the pronoun can refer to the correct referent. Thus we did not use the standard metric in coreference resolution evaluation, such as MUC (Vilain et al., 1995), or B^3 (Bagga and Baldwin, 1998). Instead, we compare the pronouns that refer to "NP", "VP" or "Sentence" in our annotation and those from the automatic tools, and calculate the precision, recall and *F*-score. Note that the results here are not comparable with those reported in prior coreference work, since we use a subset of the system output and also different evaluation metrics.

The phrase boundaries in the human annotation and the automatic systems are not always the same, hence we use some rules for evaluation purpose. For Cherrypicker and Reconcile, we allow partial match: if over half of the words in the referred phrase are contained in the human annotation, we consider it correct. Because the Stanford tool also labels the head word of each phrase, we consider it correct only if the head word is contained in the gold annotation.

Table 13 shows the pronoun coreference resolution results for the three tools. We can see that none of them performs well on our defined pronoun set in this corpus. In particular, they all have quite a low recall. The best performing system is Reconcile. We also analyze the performance on each pronoun by comparing human annotation and the results from Reconcile. Table 14 shows the number of examples that are true positive, true negative, false positive, and false negative in each pronoun. The results show that "it", "they" and "that" are the majority in all the 11 pronouns and the system fails to capture most of them (large number in false negative). These pronouns are hard to resolve for an automatic system, because they are more ambiguous than other pronouns such as "he" or "she". For example, "it" has several potential antecedent classes and linguistic functions, beyond referring to the explicitly mentioned antecedent, and it can be used nonreferentially (Foraker and McElree, 2007). It shows the limit of using automatic coreference resolution system to summarize speech corpus.

Table 14
Performance of Reconcile on each individual pronoun.

	It	They	Them	Their	He	His	Him	She	Her	This	That
True positive	24	45	8	0	8	2	2	2	3	0	0
True negative	271	70	5	11	4	1	0	0	0	49	550
False positive	89	11	4	0	0	0	0	0	0	0	5
False negative	252	213	67	42	57	4	5	13	8	22	222

The following shows some false negative examples. They are either labeled incorrectly or not labeled by the system. In the examples, the pronouns and the reference referred entities are shown in bold. We notice that for the automatic coreference system, many errors occur because of the inaccurate noun phrase identification, such as in Example1: "he" refers to "the guy", but the system labels as "Eleven". Example 2 and 3 are not labeled by all three systems.

[**Example1:**]

- *B: like I say, I don't think **the guy who's going to rob a Seven Eleven**, is going to rob a Seven Eleven whether <pron value="Eleven" category="NP" refline="59">**he**</pron> has a gun or a knife, baseball bat,*
- *A: exactly.*
- *B: or, you know, whatever,*
- *A: <pron value="Eleven" category="NP" refline="59"> **He**</pron> is going to do what <pron id=" " value="Eleven" category="NP" refline="59">**he**</pron> wants to do.*

[**Example2:**]

- *B: But, I guess I believe, I think **the N R A** has gone overboard the wrong way. You know,*
- *A: Yeah.*
- *B: **They** are saying absolutely no gun control.*

[**Example3:**]

- *B: I'm not really sure what Texas law, I think there's **a check for felonies**, on your record.*
- *A: Right.*
- *B: If the gun shop owner does **it**, -*

8. Conclusions

In this study we presented our work on opinion summarization on spontaneous conversations. We annotated a set of phone conversations from the Switchboard corpus with sentiment labels and human summarization. We adopted two approaches: the first is an unsupervised method where we incorporated dialogue structure into the traditional graph-based method; the second is a supervised approach where we use salience score, sentiment score, topic relevance, and sentence length as features. Our experimental results showed that both methods are able to improve beyond the longest sentence length baseline approach. Furthermore, given the evidence that pronouns are frequently used in conversations but ignored by general extractive summarization systems, we investigated incorporating pronoun resolution information in spontaneous conversation summarization. We collected pronoun coreference annotation, and analyzed the relation between the summaries and pronouns. We developed various features based on pronoun coreference and incorporated them in our supervised summarization system. The experimental results showed that the pronoun related features are helpful in generating better summaries. Lastly, we compared the results from three state-of-the-art automatic coreference resolution systems with human annotation, and found that they cannot perform well on the spontaneous conversation corpus. Our analysis showed that the pronouns most frequently used in conversations are those that are quite hard to resolve for the automatic systems, which showed the limit of using automatically generated pronoun resolution result in speech summarization. In our future work, we plan to explore more robust features for summarization for the ASR condition, and improve automatic pronoun resolution performance for conversations.

Acknowledgments

The authors thank the reviewers for the valuable comments, Julia Hirschberg and Ani Nenkova for discussion of the work, and Wen Wang for helping with the generation of the speech recognition output for the Switchboard data. This research is supported by NSF awards CNS-1059226 and IIS-0845484. Any opinions, findings, and conclusions or recommendations expressed are those of the author and do not necessarily reflect the views of NSF.

References

Azzam, S., Humphreys, K., Gaizauskas, R., 1999. Using coreference chains for text summarization. In: Proceedings of the Workshop on Coreference and its Applications, pp. 77–84.

Bagga, A., Baldwin, B., 1998. Algorithms for scoring coreference chains. In: Proceedings of Language Resources and Evaluation Workshop on Linguistics Coreference, pp. 563–566.

Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P., 2010. Going beyond traditional QA systems: challenges and keys in opinion question answering. In: Proceedings of COLING, pp. 27–35.

Berger, A.L., Pietra, S.A.D., Pietra, V.J.D., 1996. A maximum entropy approach to Natural Language Processing. Comput. Linguist. 22, 39–71.

Bergler, S., Witte, R., Khalife, M., Li, Z., Rudzicz, F., 2003. Using knowledge-poor coreference resolution for text summarization. In: Proceedings of DUC, pp. 85–92.

Chen, Y.-N., Metze, F., 2012. Intra-speaker topic modeling for improved multi-party meeting summarization with integrated random walk. In: Proceedings of NAACL:HLT, pp. 377–381 http://www.aclweb.org/anthology/N12-1041

Daume III, H., 2007. Frustratingly easy domain adaptation. In: Proceedings of ACL, pp. 256–263 http://www.aclweb.org/anthology/P07-1033

Erkan, G., Radev, D.R., 2004. Lexrank: graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. 22, 457–479.

Foraker, S., McElree, B., 2007. The role of prominence in pronoun resolution: active versus passive representations. J. Memory Lang. 56 (3), 357–383.

Furui, S., Kikuchi, T., Shinnaka, Y., Hori, C., 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. IEEE Trans. Audio Speech Lang. Process. 12 (4), 401–408.

Ganesan, K., Zhai, C., Han, J., 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In: Proceedings of COLING, pp. 340–348.

Garg, N., Favre, B., Reidhammer, K., Tür, D.H., 2009. Clusterrank: a graph based method for meeting summarization. In: Proceedings of Interspeech.

Gillick, D., Riedhammer, K., Favre, B., Tür, D.H., 2009. A global optimization framework for meeting summarization. In: Proceedings of ICASSP, pp. 4769–4772.

Godfrey, J.J., Holliman, E., 1997. Switchboard-1 release 2. Linguistic Data Consortium, Philadelphia.

Hayes, A., Krippendorff, K., 2007. Answering the call for a standard reliability measure for coding data. Commun. Meth. Meas. 1 (1), 77–89.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. In: Proceedings of ACM SIGKDD, pp. 168–177.

Joseph, F.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76 (5), 378–382.

Koumpis, K., Renals, S., 2005. Automatic summarization of voicemail messages using lexical and prosodic features. ACM Trans. Speech Lang. Process. 2 (1).

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D., 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In: Proceedings of CONLL: Shared Task, pp. 28–34.

Lerman, K., McDonald, R., 2009. Contrastive summarization: an experiment with consumer reviews. In: Proceedings of NAACL, pp. 113–116.

Lin, C.-Y., 2004. Rouge: a package for automatic evaluation of summaries. In: Proceedings of ACL workshop on Text Summarization Branches Out, pp. 74–81.

Lin, H., Bilmes, J., 2010. Multi-document summarization via budgeted maximization of submodular functions. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 912–920.

Lin, S.H.L., Chen, B.C., Wang, H.-m., 2009. A comparative study of probabilistic ranking models for Chinese spoken document summarization. ACM Trans. Asian Lang. Inform. Process. 8 (1), 1–23.

Lin, Z., Liu, C., Ng, H.T., Kan, M.-Y., 2012. Combining coherence models and machine translation evaluation metrics for summarization evaluation. In: Proceedings of ACL, pp. 1006–1014.

Liu, F., Liu, Y., 2008. What are meeting summaries? an analysis of human extractive summaries in meeting corpus. In: Proceedings of SIGDial, pp. 80–83.

Maloof, M.A., 2003. Learning when data sets are imbalanced and when costs are unequal and unknown. In: ICML Workshop on Learning from Imbalanced Data Sets II.

Maskey, S., Hirschberg, J., 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In: Proceedings of Interspeech.

Meng, X., Wei, F., Liu, X., Zhou, M., Li, S., Wang, H., 2012. Entity-centric topic-oriented opinion summarization in twitter. In: Proceedings of ACM SIGKDD, pp. 379–387.

Murray, G., Carenini, G., 2008. Summarizing spoken and written conversations. In: Proceedings of EMNLP, pp. 773–782.

Murray, G., Carenini, G., 2009. Detecting subjectivity in multiparty speech. In: Proceedings of Interspeech.

Murray, G., Renals, S., Carletta, J., 2005. Extractive summarization of meeting recordings. In: Proceedings of Eurospeech.

Ng, V., Dasgupta, S., Arifin, S.M.N., 2006. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In: Proceedings of COLING/ACL, pp. 611–618.
Nigam, K., 1999. Using maximum entropy for text classification. In: Proceedings of IJCAI Workshop on Machine Learning for Information Filtering, pp. 61–67.
Nishikawa, H., Hasegawa, T., Matsuo, Y., Kikui, G., 2010. Opinion summarization with integer linear programming formulation for sentence extraction and ordering. In: Proceedings of COLING, pp. 910–918.
Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The PageRank citation ranking: bringing order to the web.
Pang, B., Lee, L., 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of ACL, pp. 271–278.
Paul, M., Zhai, C., Girju, R., 2010. Summarizing contrastive viewpoints in opinionated text. In: Proceedings of EMNLP, pp. 66–76.
Popescu, A.-M., Etzioni, O., 2007. Extracting product features and opinions from reviews., pp. 9–28.
Raaijmakers, S., Truong, K., Wilson, T., 2008. Multimodal subjectivity analysis of multiparty conversation. In: Proceedings of EMNLP, pp. 466–474.
Rahman, A., Ng, V., 2009. Supervised models for coreference resolution. In: Proceedings of EMNLP, pp. 968–977.
Shen, D., Sun, J.-T., Li, H., Yang, Q., Chen, Z., 2007. Document summarization using conditional random fields. In: Proceedings of IJCAI, pp. 2862–2867.
Steinberger, J., Kabadjov, M.A., Poesio, M., Sanchez-Graillet, O., 2005. Improving LSA-based summarization with anaphora resolution. In: Proceedings of HLT/EMNLP, pp. 1–8.
Stoyanov, V., Cardie, C., 2006. Partially supervised coreference resolution for opinion summarization through structured rule learning. In: Proceedings of EMNLP, pp. 336–344.
Stoyanov, V., Cardie, C., Gilbert, N., Riloff, E., Buttler, D., Hysom, D., 2010. Coreference resolution with Reconcile. In: Proceedings of ACL, pp. 156–161.
Stoyanov, V., Cardie, C., Wiebe, J., 2005. Multi-perspective question answering using the OpQA corpus. In: Proceedings of EMNLP/HLT, pp. 923–930.
Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L., 1995. A model-theoretic coreference scoring scheme. In: Proceedings of Message understanding, pp. 45–52, http://dx.doi.org/10.3115/1072399.1072405.
Wang, D., Liu, Y., 2011. A pilot study of opinion summarization in conversations. In: Proceedings of ACL.
Wiebe, J., Riloff, E., 2005. Creating subjective and objective sentence classifiers from unannotated texts. In: Proceedings of CICLing, pp. 486–497.
Wilson, T., 2008. Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states (Ph.D. thesis). University of Pittsburgh.
Wilson, T., Wiebe, J., 2003. Annotating opinions in the world press. In: Proceedings of SIGDial, pp. 13–22.
Wong, K.-F., Wu, M., Li, W., 2008. Extractive summarization using supervised and semi-supervised learning. In: Proceedings of COLING, pp. 985–992.
Xie, S., Hakkani-Tur, D., Favre, B., Liu, Y.,2009. Integrating prosodic features in extractive meeting summarization. In: Proceedings of IEEE Workshop on ASRU. IEEE, pp. 387–391.
Xie, S., Liu, Y., 2010. Improving supervised learning for meeting summarization using sampling and regression. Comput. Speech Lang. 24 (3), 495–514.
Zechner, K., 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. Comput. Linguist. 28 (4), 447–485.
Zhang, J.J., Chan, H.Y., Fung, P., 2007. Improving lecture speech summarization using rhetorical information. In: Proceedings of IEEE Workshop on ASRU, pp. 195–200.
Zhao, L., Wu, L., Huang, X., 2009. Using query expansion in graph-based approach for query-focused multi-document summarization. J. Inform. Process. Manag. 45 (1), 35–41.
Zhu, X., Penn, G., 2006. Summarization of spontaneous conversations. In: Proceedings of Interspeech, pp. 1531–1534.