EVALUATING THE EFFECTIVENESS OF FEATURES AND SAMPLING IN EXTRACTIVE MEETING SUMMARIZATION

Shasha Xie, Yang Liu

Department of Computer Science
The University of Texas at Dallas
{shasha, yangl}@hlt.utdallas.edu

Hui Lin

Department of Electrical Engineering
University of Washington
hlin@ee.washington.edu

ABSTRACT

Feature-based approaches are widely used in the task of extractive meeting summarization. In this paper, we analyze and evaluate the effectiveness of different types of features using Forward Feature Selection in an SVM classifier. In addition to features used in prior studies, we introduce topic related features and demonstrate that these features are helpful for meeting summarization. We also propose a new way to resample the sentences based on their salience scores for model training and testing. The experimental results on both the human transcripts and recognition output, evaluated by the ROUGE summarization metrics, show that feature selection and data resampling help improve the system performance.

Index Terms— meeting summarization, forward feature selection, resampling, TFIDF

1. INTRODUCTION

Extractive meeting summarization is a useful tool to facilitate users to browse the meeting recordings. Compared to summarization of written text and other speech genre, there are many challenges in the meeting domain because of its style, such as the presence of disfluencies, multiple speakers, the high recognition error rate, and less coherence in the context. Feature-based approaches have been widely used for speech summarization recently, such as Support Vector Machines (SVM) [1], Hidden Markov Models (HMM) [2], Conditional Random Fields (CRF) [3], and maximum entropy model [4]. In these approaches, extractive speech summarization is considered as a binary classification problem, that is, for each sentence, a decision is made whether to select it into the summary or not.

The features that are commonly used for speech summarization can be broadly classified as lexical, structural, discourse, and prosodic features. A few studies were performed to analyze the effectiveness of different types of features for summarization in the domain of broadcast news and lectures [2, 5]. However, less work has examined these features for the task of meeting summarization. We believe this analysis is important because meeting speech is significantly different from other speech genre in that it is more spontaneous and not well-structured. In addition to the widely used features, we introduce topic related features in this study. The meeting transcripts are often very long (e.g., corresponding to an hour meeting), and can be divided into different topics. For each word and sentence, we thus derive features based on their distribution in different topics to help determine their importance. In this paper, we evaluate the contributions of various features using an SVM classifier and Forward Feature Selection on the ICSI meeting corpus. We show that using a subset of features outperforms all the features, and that the effective features are from different categories, including the topic related features we added.

Since we treat the summarization task as a binary classification problem, and the positive class (summary sentences) is the minority class, we also consider the imbalance problem in this study. In the ICSI meeting corpus, the average positive percentage is 6.62%. This imbalanced data often raises problems for a classifier. For the meeting summarization task, this problem has never been investigated previously. In this paper, we propose a data resampling approach to improve classification performance. Our goal is to change the ratio of the positive instances to the negative ones by re-selecting the training samples. The criteria we use to select samples is their salience scores (TFIDF values of all the words in a sentence), which in general give higher weights to positive instances and lower weights to negative instances. We only preserve a certain percentage of sentences in training and testing. Our experimental results have shown that this method yields improvement, especially when using speech recognition output.

2. FEATURES FOR MEETING SUMMARIZATION

2.1. Features

We start our work by extracting a variety of features, including those that have been used for text and speech summarization in previous work and the new ones using topic information.

A: Lexical Features[1]

Table 1 lists the lexical features. The first part includes the sentence length and the number of words in each sentence after removing the stop words. "Unigram" and "Bigram" are the number of frequent words and bigrams in the sentence based on the list we automatically generated using a development set [3]. Finally previous work has shown that the first appearing nouns and pronouns in a sentence provide important new information [6], therefore we use features to represent the number of nouns or pronouns that appear for the first time in a sentence.

B: Structural and Discourse Features

The structural and discourse features are described in Table 2. Cosine similarity between two text segments (D_1 and D_2) is:

$$sim(D_1, D_2) = \frac{\sum_i t_{1i} t_{2i}}{\sqrt{\sum_i t_{1i}^2} \times \sqrt{\sum_i t_{2i}^2}} \quad (1)$$

where t_i is the term weight for a word w_i, for which we use the TF-IDF (term frequency, inverse document frequency) value. We

[1]Note that the names we use for different categories may not be perfect. But we expect that by listing all the features along with their description, it is clear about all the features used in this study.

Feature	Feature Description
Len I, II, III	previous, current and next sentence length
Num I, II, III	# of words in previous, current, and next sentence respectively (removing stopwords)
Unigram	# of frequent words
Bigram	# of frequent bigrams
Noun	# of first appearing nouns
Pronoun	# of first appearing pronouns

Table 1. List of lexical features

also derive various TF and IDF related features (e.g., max, mean, sum) for a sentence. Features "Speaker" and "Same_as_prev" are used to represent the speaker information. According to the sentence length, we automatically select the main speakers for each meeting. Each sentence is labeled whether it is said by the main speaker, and whether the speaker is the same as the previous one. To capture how term usage varies across speakers in a given meeting, we adopt the feature SUIDF introduced in [7]. The hypothesis for this feature is that more informative words are used with varying frequencies among different meeting participants, and less informative words are used rather consistently by different speakers.

Feature	Feature Description
Cosine	cosine similarity between each sentence to the whole document
TF I, II, III	Mean, Max, and Sum of TF
IDF I, II, III	Mean, Max, and Sum of IDF
TFIDF I, II, III	Mean, Max, and Sum of TF*IDF
Speaker	main speaker or not
Same_as_prev	same as the previous speaker or not
SUIDF	Mean, Max, and Sum of SUIDF

Table 2. List of structural and discourse features

C: Topic-related Features

Even though the meeting transcripts are not as organized as broadcast news speech (which generally consists of better story segments), they can still be divided into several parts, each with its own topic. We believe that topic segmentation contains useful information for the summarization of a meeting document. To better capture the characteristics of different topics in a meeting, several topic related features are introduced in our study.

The topic related features we use are based on the so-called topic term frequency (TTF) and inverse topic frequency (ITF), both of which are calculated on a topic basis for each meeting transcript. The TTF is just the term frequency within a topic, and the ITF values are computed as:

$$ITF(w_i) = log(NT/NT_i) \quad (2)$$

where NT_i is the number of topics containing word w_i within a meeting, and NT is the total number of topic segments in this meeting. Note that ITF values are estimated for each meeting, whereas the earlier IDF values are calculated based on the entire corpus. Our hypothesis is that this meeting specific ITF might be more indicative of a specific topic in this meeting. Table 3 lists the topic related features.

2.2. Forward Feature Selection

In order to analyze the importance of different features described above, we use Forward Feature Selection (FFS) to rank each feature [8], as shown in Algorithm 1.

Feature	Feature Description
ITF I, II, III	Mean, Max, and Sum of ITF
TTFITF I, II, III	Mean, Max, and Sum of TTF*ITF

Table 3. List of topic related features

Algorithm 1 Algorithm for Forward Feature Selection

Let $P = \emptyset$ **be the current set of selected features**
Let Q **be the full set of features**
while size of P is smaller than a given constant **do**
  (a) **for each** $v \in Q$
  Set $P' \leftarrow P \cup \{v\}$
  Train the model with P' and evaluate on the dev set
  (b) **Set** $P \leftarrow P \cup \{v^*\}$
  where v^* is the best feature obtained in step (a)
  (c) **Set** $Q \leftarrow Q \backslash \{v^*\}$
  (d) **Record the validation performance obtained with the current** P
end while

3. SUMMARIZATION APPROACH

Support Vector Machine (SVM) is the classifier we use in our experiments. For each sentence in the meeting transcript, we predict its confidence scores of being included into the summary. Then the summary for the meeting is constructed by selecting a predefined percentage of the sentences/words with the highest scores.

We propose to use a resampling technique in the classification approach to select a subset of the sentences for model training and testing. The selection is done based on the salience score of each sentence. We investigate two methods for computing the sentence weight: one is based on the TFIDF value of the words in the sentence; the other is the cosine similarity of the sentence and the entire document. In genreal, both of these two methods give higher scores to summary sentences than non-summary sentences. We will evaluate the impact on summarization using different resampling rate (e.g., percentage of the sentences preserved based on their weights).

The benefit of this resampling is two fold. First, this helps address the imbalanced data problem in training. In each meeting document, only a small percentage of the sentences are labeled as summary sentence, resulting in a very low ratio of positive versus negative instances for classifier training. Our sentence selection can preserve most of the positive instances and remove negative instances (as verified later in Section 4.2.2), thus increasing the positive to negative ratio for classifier training. Second, during testing, this sampling approach only keeps those sentences with a high weight and reduces the number of candidate sentences. Since most of the sentences ignored are non-summary sentences, this does not have a negative impact, but rather allowing the model to focus on the more likely candidate sentences.

4. EXPERIMENTS

4.1. Corpus and Experimental Setup

We use the ICSI meeting corpus [9], which contains 75 recordings from natural meetings. Each meeting is about an hour long. These meetings have been transcribed, and the annotated dialog acts (DA) [10] are used as the sentence units for extractive meeting summarization. The meeting corpus has also been annotated with topic segmentation and extractive summaries [11]. The ASR output is obtained from a state-of-the-art SRI conversational telephone speech

system [12], with a word error rate of about 38.2% on the entire corpus. We align the human annotated DA boundaries to the ASR words to obtain the DAs for the ASR transcripts. Similarly, topic boundaries for the ASR output are obtained by aligning the human annotated topic segments to the ASR words.

We use the same 6 meetings as in [13] to form the test set, and the other 69 meetings as the training set. Furthermore, 6 meetings are randomly selected from the training set as the development set, then the rest is used to compose the training corpus for the SVM classifier. The development set is used to select the top features using FFS, and to decide the percentage of the sentences used in re-sampling for training and test. We use the 69 training meetings to calculate the IDF values. For both the human transcripts and ASR output, we split each of the 69 training meetings into multiple topics, and then use these new "documents" to calculate the IDF values.

To evaluate summarization performance, we use ROUGE [14], which has been used in previous studies of speech summarization [1, 13, 15]. ROUGE compares the system generated summary with the reference summaries, and measures different matches, such as N-gram, longest common sequence, and skip bigrams. It can accept multiple reference summaries. For the 6 test meetings, we use 3 human annotated summaries as the reference.

4.2. Experimental Results and Discussion

4.2.1. Results Using Forward Feature Selection

For FFS using the SVM classifier, instead of using classification accuracy or error rate, we use the ROUGE score (F-measure of R-1, unigram match) as the metric, since that is our ultimate performance measure. We train an SVM model, and predict each sentence's confidence score of being in the summary on the development set. The summary is extracted by selecting the sentences with the highest probabilities, with a word compression ratio of 18%.

The top features selected incrementally by FFS are listed in Table 4, along with the ROUGE-1 F-measure. The ROUGE score when using all the features is also shown in the table for a comparison. It suggests that a subset of the features can outperform using all the features, therefore feature selection is important for this task. The 6 selected features are from different categories, including lexical features (*Len II*, *Bigram*, *Noun*), structural features (*IDF II*, *TF II*), and topic related features (*ITF II*). The sentence length has been proved to be a very important feature, and a very competitive baseline by previous work [16]. *Bigram* and *Noun* are also included in the top features, which shows that the lexical cues are also very predictive in the domain of meeting summarization. The topic related feature, *ITF II*, that we have introduced is also selected and its addition yields noticeable gain.

Feature	Feature Description	ROUGE
LenII	current sentence length	0.678
+ Bigram	most frequent bigram	0.683
+ IDF II	Max of IDF	0.686
+ Noun	first appearing noun	0.687
+ TF II	max of TF	0.685
+ ITF II	Max of ITF	0.692
All features		0.669

Table 4. Forward Feature Selection results.

4.2.2. Results Using Re-sampling

In order to verify that after removing the sentences with lower weights, the remaining samples still include most of the positive samples (i.e., the training data are more balanced), we calculate the average coverage of all the original positive sentences and the percentage of positive instances after re-sampling for different sampling rate on the training data. Results are illustrated in Figure 1. We can see that the top 50% sentences can preserve 94.3% of the positive sentences when using TFIDF scores as the selection criteria. The average positive percentage using resampling is much higher than that in the original data (12.47% vs. 6.62%). Figure 1 also shows that TFIDF scores outperform cosine similarity in terms of the coverage of positive samples or the percentage of positive sentences, therefore we use TFIDF-based resampling in the following experiments.

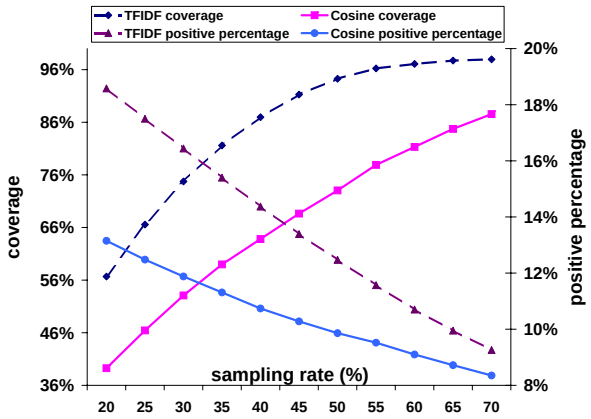

Fig. 1. Coverage of the original positive samples (left Y-axis) and the percentage of positive samples in the selected data (right Y-axis) using TFIDF and cosine scores as selection criteria for different re-sampling rates.

We use the resampled data to train the SVM model, and use the same sampling rate during testing. We have evaluated different sampling rates and word compression ratios on both the human transcripts and ASR output of the development set. The ROUGE-1 F-measure results are showed in Figure 2 for human transcripts and Figure 3 for ASR output respectively, with a sampling rate of 25%, 40%, and 50%. All of the results here are based on the 6 features selected in Sec 4.2.1. For comparison, we also include the results without re-sampling as the baseline. On human transcripts, re-sampling does not improve the performance. However, on ASR output, there is a gain – for the sampling rate of 40%, the ROUGE scores on all the word compression ratio are consistently better than the baseline; and with a sampling rate of 25% the word compression ratio of 17%, we obtain the best ROUGE score, 0.66.

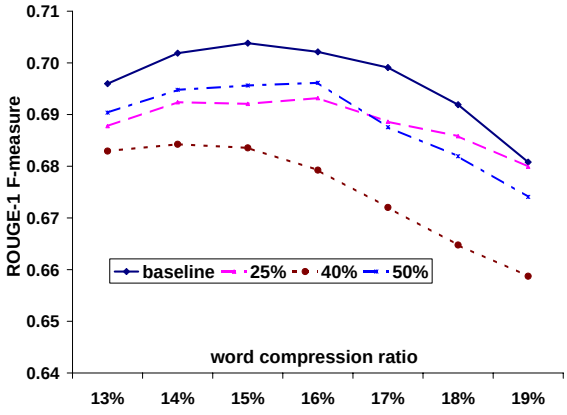

Fig. 2. Results on dev set using human transcripts.

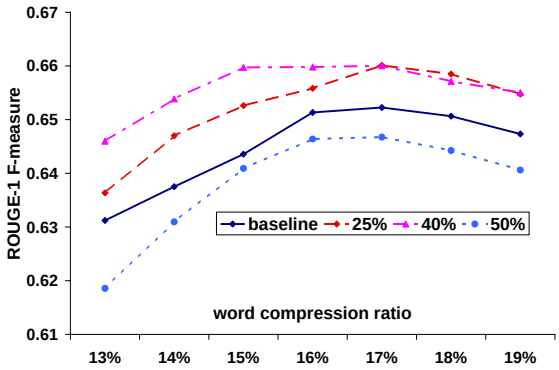

Fig. 3. Results on dev set using ASR output.

4.2.3. Results on ASR Test Set

The results on the test set using ASR output are shown in Figure 4 for various conditions: baseline using all the features, the 6 selected features without resampling, and the 6 selected features with resampling for two resampling rates (chosen based on the results from the dev set). We can see that using all the features as described in Section 2 yields the lowest ROUGE scores. The 6 features selected from human transcripts work well on ASR output. The two different sampling rates give similar results, both better than without re-sampling.

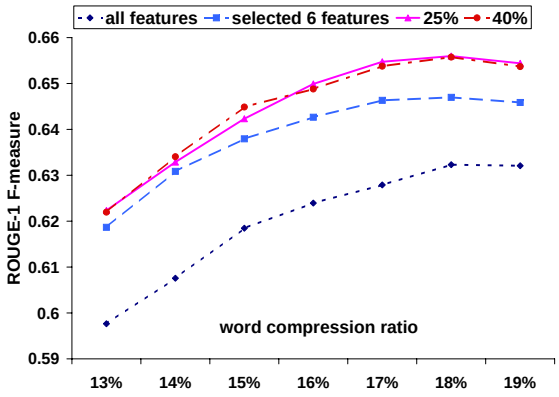

Fig. 4. Results on test set for ASR output.

5. CONCLUSION AND FUTURE WORK

In this paper, we used the SVM model for automatic extractive meeting summarization. We have extracted different types of features, including the lexical, structural, discourse, and topic related features. The effectiveness of these features for the task of meeting summarization has been evaluated through Forward Feature Selection. We were able to select a subset of features that outperform using all the features. The topic related features that we introduced in this paper also proved to be helpful. In addition, we investigated data resampling techniques for classifier training and testing. It can effectively deal with the imbalanced data problem for summarization. Using TFIDF-based selection provides a better coverage of the positive instances and is superior to using cosine similarity based weights. Our experimental results on the blind test set using ASR output proved that the feature selection and re-sampling method can significantly improve the system performance.

The feature selection was conducted on the human transcripts, we will evaluate the effectiveness of these features on ASR output in our future work. We will also add prosodic features and evaluate their impact on this task. In addition, the meeting summarization task is considered as a binary classification problem in this paper. We will investigate using regression techniques for this task. Finally, we will incorporate automatic dialog act segmentation and topic segmentation in our future work.

6. ACKNOWLEDGMENTS

The authors thank University of Edinburgh for sharing the annotation on the ICSI meeting corpus. This research is supported by NSF award IIS-0714132.

7. REFERENCES

[1] Justin Jian Zhang, Ho Yin Chan, and Pascale Fung, "Improving lecture speech summarization using rhetorical information," in *ASRU*, 2007.

[2] Sameer Maskey and Julia Hirschberg, "Summarizing speech without text using Hidden Markov Models," in *HLT-NAACL*, 2006.

[3] Michel Galley, "A skip-chain conditional random field for ranking meeting utterances by importance," in *EMNLP*, 2006.

[4] Anne Hendrik Buist, Wessel Kraaij, and Stephan Raaijmakers, "Automatic summarization of meeting data: A feasibility study," in *the 15th CLIN conference*, 2005.

[5] Jian Zhang and Pascale Fung, "Speech summarization without lexical features for mandarin broadcast news," in *HLT-ACL*, 2007.

[6] Sameer Maskey and Julia Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *Interspeech*, 2005.

[7] Gabriel Murray and Steve Renals, "Term-weighting for summarization of multi-party spoken dialogues," in *MLMI*, 2007.

[8] Isabelle Guyon and André Elisseeff, "An introduction to variable and feature selection," in *Journal of Machine Learning Research*, 2003, vol. III, pp. 1157–1182.

[9] Adam Janin, Don Baron, Jane Edwards, and Dan Ellis et al., "The ICSI meeting corpus," in *ICASSP*, 2003.

[10] Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *the 5th SIGDIAL Workshop*, 2004.

[11] Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore, "Evaluating automatic summaries of meeting recordings," in *the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, 2005.

[12] Qifeng Zhu, Andreas Stolcke, Barry Chen, and Nelson Morgan, "Using MLP features in SRI's conversational speech recognition system," in *Interspeech*, 2005.

[13] Gabriel Murray, Steve Renals, and Jean Carletta, "Extractive summarization of meeting recordings," in *Interspeech*, 2005.

[14] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," in *the Workshop on Text Summarization Branches Out*, 2004.

[15] Xiaodan Zhu and Gerald Penn, "Summarization of spontaneous conversations," in *Interspeech*, 2006.

[16] Gerald Penn and Xiaodan Zhu, "A critical reassessment of evaluation baselines for speech summarization," in *ACL-HLT*, 2008.