Speech Summarization Without Lexical Features for Mandarin Broadcast News

Jian Zhang
Human Language Technology Center
Electronic and Computer Engineering
University of Science and Technology
Clear Water Bay,Hong Kong
zjustin@ust.hk

Pascale Fung
Human Language Technology Center
Electronic and Computer Engineering
University of Science and Technology
Clear Water Bay,Hong Kong
pascale@ee.ust.hk

Abstract

We present the first known empirical study on speech summarization without lexical features for Mandarin broadcast news. We evaluate acoustic, lexical and structural features as predictors of summary sentences. We find that the summarizer yields good performance at the average F-measure of 0.5646 even by using the combination of acoustic and structural features alone, which are independent of lexical features. In addition, we show that structural features are superior to lexical features and our summarizer performs surprisingly well at the average F-measure of 0.3914 by using only acoustic features. These findings enable us to summarize speech without placing a stringent demand on speech recognition accuracy.

1 Introduction

Speech summarization, a technique of extracting key segments that convey the main content from a spoken document or audio document, has become a new area of study in the last few years. There has been much significant progress made in speech summarization for English or Japanese text and audio sources (Hori and Furui, 2003; Inoue et al., 2004; Koumpis and Renals, 2005; Maskey and Hirschberg, 2003; Maskey and Hirschberg, 2005). Some research efforts have focused on summarizing Mandarin sources (Chen et al., 2006; Huang et al., 2005), which are dependent on lexical features. Considering the difficulty in obtaining high quality transcriptions, some researchers proposed speech summarization systems with non-lexical features (Inoue et al., 2004; Koumpis and Renals, 2005; Maskey and Hirschberg, 2003; Maskey and Hirschberg, 2006). However, there does not exist any empirical study on speech summarization without lexical features for Mandarin Chinese sources. In this paper, we construct our summarizer with acoustic and structural features, which are independent of lexical features, and compare acoustic and structural features against lexical features as predictors of summary sentences.

In Section 2 we review previous work on broadcast news summarization. We describe the Mandarin broadcast news corpus on which our system operates in Section 3. In Section 4 we describe our summarizer and these features used in experiments. We set up our experiments and evaluate the results in Section 5, followed by our conclusion in Section 6.

2 Previous Work

There have been many research efforts on speech summarization. Some methods dependent on lexical features are presented (Inoue et al., 2004; Chen et al., 2006; Huang et al., 2005). (Inoue et al., 2004) uses statistical methods to identify words to include in a summary, based on linguistic and acoustic/prosodic features of the Japanese broadcast news transcriptions; while (Chen et al., 2006) proposes the use of probabilistic latent topical information for extractive summarization of Mandarin spoken documents. (Huang et al., 2005) presents Mandarin spo-

Proceedings of NAACL HLT 2007, Companion Volume, pages 213–216,
Rochester, NY, April 2007.

ken document summarization scheme using acoustic, prosodic, and semantic information. Alternatively, some methods which are independent of lexical features are presented (Maskey and Hirschberg, 2003; Maskey and Hirschberg, 2006). (Maskey and Hirschberg, 2003) extracts structural information from audio documents to help summarization. (Maskey and Hirschberg, 2006) focuses on how to use acoustic information alone to help predict sentences to be included in a summary and shows a novel way of using continuous HMMs for summarizing speech documents without transcriptions.

It is advantageous to build speech summarization models without using lexical features: we can summarize speech data without placing a stringent demand on the speech recognition accuracy. In this paper, we propose one such model on Mandarin broadcast news and compare the effectiveness of acoustic and structural features against lexical features as predictors of summary sentences.

3 The Corpus and Manual Summaries

We use a portion of the 1997 Hub4 Mandarin corpus available via LDC as experiment data. The related audio data were recorded from China Central Television(CCTV) International News programs. They include 23-day broadcast from 14th January, 1997 to 21st April, 1997, which contain 593 stories and weather forecasts. Each broadcast lasts approximately 32 minutes, and has been hand-segmented into speaker turns. For evaluation, we manually annotated these broadcast news, and extracted segments as reference summaries. We divide these broadcast news stories into 3 types: one-turn news, weather forecast, and several-turns news. The content of each several-turn news is presented by more than one reporter, and sometimes interviewees. We evaluate our summarizer on the several-turns news corpus. The corpus has 347 stories which contain 4748 sentences in total.

4 Features and Methodology

4.1 Acoustic/Prosodic Features

Acoustic/prosodic features in speech summarization system are usually extracted from audio data. Researchers commonly use acoustic/prosodic variation – changes in pitch, intensity, speaking rate – and duration of pause for tagging the important contents of their speeches (Hirschberg, 2002). We also use these features for predicting summary sentences on Mandarin broadcast news.

Our acoustic feature set contains thirteen features: ***DurationI***, ***DurationII***, ***SpeakingRate***, ***F0I***, ***F0II***, ***F0III***, ***F0IV***, ***F0V***, ***EI***, ***EII***, ***EIII***, ***EIV*** and ***EV***. *DurationI* is the sentence duration. *DurationII* is the average phoneme duration. General phonetic studies consider that the speaking rate of sentence is reflected in syllable duration. So we use average syllable duration for representing *SpeakingRate*. *F0I* is F0's minimum value. *F0II* is F0's maximum value. *F0III* equals to the difference between *F0II* and *F0I*. *F0IV* is the mean of F0. *F0V* is F0 slope. *EI* is minimum energy value. *EII* is maximum energy value. *EIII* equals to the difference between *EII* and *EI*. *EIV* is the mean of energy value. *EV* is energy slope. We calculate *DurationI* from the annotated manual transcriptions that align the audio documents. We then obtain *DurationII* and *SpeakingRate* by phonetic forced alignment. Next we extract F0 features and energy features from audio data by using Praat (Boersma and Weenink, 1996).

4.2 Structural Features

Each broadcast news of the 1997 Hub4 Mandarin corpus has similar structure, which starts with an anchor, followed by the formal report of the story by other reporters or interviewees.

Our structural feature set consists of 4 features: ***Position***, ***TurnI***, ***TurnII*** and ***TurnIII***. *Position* is defined as follows: one news has *k* sentences, then we set $(1-(0/k))$ as *Position* value of the first sentence in the news, and set $(1-((i-1)/k))$ as *Position* value of the i^{th} sentence. *TurnI* is defined as follows: one news has *m* turns, then we set $(1-(0/m))$ as *TurnI* value of the sentences which belong to the first turn's content, and set $(1-((j-1)/m))$ as *TurnI* values of the sentences which belong to the j^{th} turn's content. *TurnII* is the previous turn's *TurnI* value. *TurnIII* is the next turn's *TurnI* value.

4.3 Reference Lexical Features

Most methods for text summarization mainly utilize lexical features. We are interested in investigating the role of lexical features in comparison to other features. All reference lexical features are extracted

from the manual transcriptions.

Our lexical feature set contains eight features: ***LenI***, ***LenII***, ***LenIII***, ***NEI***, ***NEII***, ***NEIII***, ***TFIDF*** and ***Cosine***. For a sentence, we set the number of words in the sentence as *LenI* value. *LenII* is the previous sentence's *LenI* value. *LenIII* is the next sentence's *LenI* value. For a sentence, we set the number of Named Entities in the sentence as the *NEI* value. We define the number of Named Entities which appear in the sentence at the first time in a news as *NEII* value. *NEIII* value equals to the ratio of the number of unique Named Entities to the number of all Named Entities.

TFIDF is the product of *tf* and *idf*. *tf* is the fraction: the numerator is the number of occurrences of the considered word and the denominator is the number of occurrences of all words in a story. *idf* is the logarithm of the fraction: the numerator is the total number of sentences in the considered news and the denominator is the number of sentences where the considered word appears. *Cosine* means cosine similarity measure between two sentence vectors.

4.4 Summarizer

Our summarizer contains the preprocessing stage and the estimating stage. The preprocessing stage extracts features and normalizes all features by equation (1).

$$N_j = \frac{w_j - mean(w_j)}{dev(w_j)} \tag{1}$$

Here, w_j is the original value of feature j which is used to describe sentence i; $mean(w_j)$ is the mean value of feature j in our training set or test set; $dev(w_j)$ is the standard deviation value of feature j in our training set or test set.

The estimating stage predicts whether each sentence of the broadcast news is in a summary or not. We use Radial Basis Function(RBF) kernel for constructing SVM classifier as our estimator referring to LIBSVM (Chang and Lin, 2001), which is a library for support vector machines.

5 Experiments and Evaluation

We use the several-turn news corpus, described in Section 3, in our experiments. We use 70% of the corpus consisting of 3294 sentences as training set and the remaining 1454 sentences as held-out test set, upon which our summarizer is tested.

Table 1: Feature set Evaluation by F-measure

Feature Set	SR10%	SR15%	SR20%	Ave
Ac+St+Le	.5961	.546	.5544	.5655
Ac+St	**.5888**	**.5489**	**.5562**	**.5646**
St	.5951	.5616	.537	.5645
Le	.5175	.5219	.5329	.5241
Ac	.3068	.4092	.4582	.3914
Baseline	.21	.32	.43	.32

Ac: Acoustic; St: Structural; Le: Lexical

We measure our summarizer's performance by precision, recall, and F-measure (Jing et al., 1998). We explain these metrics as follows:

$$\text{precision} = \frac{S_{man} \bigcap S_{sum}}{S_{sum}} \tag{2}$$

$$\text{recall} = \frac{S_{man} \bigcap S_{sum}}{S_{man}} \tag{3}$$

$$\text{F-Measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

In equation (2), (3) and (4), S_{man} is the sentence set of manual summaries or reference summaries; S_{sum} is the sentence set of predicted summaries provided by our summarizer.

We have three versions of reference summaries based on summarization ratio(SR): 10%, 15% and 20% respectively. So we build three baselines referring to different versions of reference summaries. When using SR 10% summaries, we build the baselines by choosing the first 10% of sentences from each story. Our baseline results in F-measure score are given in Table 1.

We perform three sets of experiments with different summarization ratios.

By using acoustic and structural features alone, the summarizer produces the same performance as by using all features. We can find the evidence from Table 1 and Figure 1. On average, the combination of acoustic and structural features yields good performance: F-measure of 0.5646, 24.46% higher than the baseline, only 0.09% lower than the average F-measure produced by using all features. This finding makes it possible to summarize speech without

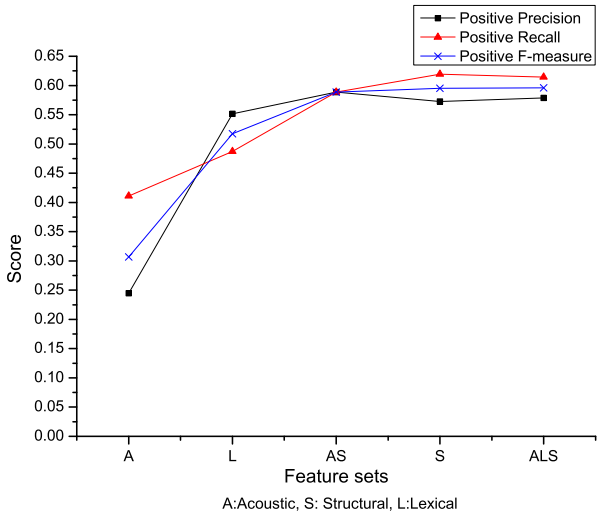

Figure 1: Performance comparison on SR10%

placing a stringent demand on the speech recognition accuracy.

In the same Mandarin broadcast program, the distribution and flow of summary sentences are relatively consistent. Therefore, compared with speech summarization on English sources, we can achieve the different finding that structural features play a key role in speech summarization for Mandarin broadcast news. Table 1 shows the evidence. On average, structural features are superior to lexical features: F-measure of 0.5645, 24.45% higher than the baseline and 4,04% higher than the average F-measure produced by using lexical features.

Another conclusion we can draw from Table 1 is that acoustic features are important for speech summarization on Mandarin broadcast news. On average, even by using acoustic features alone our summarizer yields competitive result: F-measure of 0.3914, 7.14% higher than the baseline. The similar conclusion also holds for speech summarization on English sources (Maskey and Hirschberg, 2006).

6 Conclusion

In this paper, we have presented the results of an empirical study on speech summarization for Mandarin broadcast news. From these results, we found that by using acoustic and structural features alone, the summarizer produces good performance: average F-measure of 0.5646, the same as by using all features. We also found that structural features make more important contribution than lexical features to speech summarization because of the relatively consistent distribution and flow of summary sentences in the same Mandarin broadcast program. Moreover, we have shown that our summarizer performed surprisingly well by using only acoustic features: average F-measure of 0.3914, 7.14% higher than the baseline. These findings also suggest that high quality speech summarization can be achieved without stringent requirement on speech recognition accuracy.

References

P. Boersma and D. Weenink. 1996. Praat, a system for doing phonetics by computer, version 3.4. *Institute of Phonetic Sciences of the University of Amsterdam, Report*, 132:182.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*.

B. Chen, Y.M. Yeh, Y.M. Huang, and Y.T. Chen. 2006. Chinese Spoken Document Summarization Using Probabilistic Latent Topical Information. *Proc. ICASSP*.

J. Hirschberg. 2002. Communication and prosody: Functional aspects of prosody. *Speech Communication*, 36(1):31–43.

C. Hori and S. Furui. 2003. A new approach to automatic speech summarization. *Multimedia, IEEE Transactions on*, 5(3):368–378.

C.L. Huang, C.H. Hsieh, and C.H. Wu. 2005. Spoken Document Summarization Using Acoustic, Prosodic and Semantic Information. *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 434–437.

A. Inoue, T. Mikami, and Y. Yamashita. 2004. Improvement of Speech Summarization Using Prosodic Information. *Proc. of Speech Prosody*.

H. Jing, R. Barzilay, K. McKeown, and M. Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. *AAAI Symposium on Intelligent Summarization*.

K. Koumpis and S. Renals. 2005. Automatic summarization of voicemail messages using lexical and prosodic features. *ACM Transactions on Speech and Language Processing (TSLP)*, 2(1):1–24.

S. Maskey and J. Hirschberg. 2003. Automatic summarization of broadcast news using structural features. *Proceedings of Eurospeech 2003*.

S. Maskey and J. Hirschberg. 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. *Interspeech 2005 (Eurospeech)*.

S. Maskey and J. Hirschberg. 2006. Summarizing Speech Without Text Using Hidden Markov Models. *Proc. NAACL*.