http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 13 (7): 1447-1450, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.1447.1450

Rhetorical-state SVM for Lecture Speech Summarization

Jian Zhang, Huaqiang Yuan and Xiaoheng Pan
Engineering and Technology Institute, Dongguan University of Technology, China

Abstract: A novel non-generative probabilistic framework is proposed for extractive lecture speech summarization. Rhetorical structure hidden in lecture speech is one of the most underutilized characteristics. Rhetorical State Support Vector Machine (RSSVM) is proposed for automatically decoding rhetorical structure in speech and summarizing the speech data. RSSVM gives a 57.2% ROUGE-L F-measure, a 5.6% absolute increase in lecture speech summarization performance compared with the baseline system without using rhetorical information. It also outperforms the extractive summarization directly using the rhetorical structure produced by Rhetorical-State Hidden Markov Models.

Key words: Support vector machine, rhetorical structure, speech summarization

INTRODUCTION

Automatic speech summarization is a core technology in the spoken document understanding and organization system. It is the process of digesting source speech data and producing understandable segments (in speech or transcription text form) that convey the most informative or relevant information as a substitute of the original speech. Speech summarization is young and under-exploited compared with the research field of text summarization. Unlike summarizing the written documents, summarizing spoken documents produced by automatic speech recognition system, often has a big challenge that is the lack of easily discernible structures in speech. Fonts, sentence/paragraph boundaries, title/subtitles and so on can be found and are helpful to describe the underlying meaning in the written documents. However, those kinds of structure clues do not exist in the speech data.

Acoustic and linguistic features were used for building extractive speech summarizers (Chen *et al*., 2006; Maskey and Hirschberg, 2005; Inoue *et al*., 2004; Murray *et al*., 2005). However, those summarizers all ignore rhetorical structure which exists in the spoken documents and the relevant speech data. Some researchers (Hori *et al*., 2002; Maskey and Hirschberg, 2003; Hirohata *et al*., 2005; Zhang *et al*., 2007) have suggested that rhetorical structure exists also in the spoken documents and relevant speech data and efficient representation of this information can be helpful to summarization task.

Fung *et al*. (2008) proposed Rhetorical-State Hidden Markov Models (RSHMM) for representing rhetorical structure existing in the lecture speech and RSHMM-enhanced probabilistic framework for extractive summarization. Here, this study further develops the probabilistic framework for improving summarization performance.

EXTRACTING RHETORICAL CHARACTERISTICS OF LECTURE SPEECH

Acoustic and linguistic 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). This study also investigates these features for their efficiency in predicting summary sentences on lecture speech data.

The acoustic feature set contains thirteen features: DurationI, SpeakingRate, F0I, F0II, F0III, F0IV, F0V, EI, EII, EIII, EIV and EV. These features are listed in Table 1.

Table 1: Acoustic/prosodic features and feature descriptions

Feature name	Feature description
Duration I	Time duration of sentence
Speaking rate	Average syllable duration
F0I and F0II	F0's minimum/maximum value
F0III	The difference between F0II and F0I
F0IV	Mean of F0 value
F0V	F0 slope
EI and EII	Minimum/maximum energy value
EIII	The difference between EII and EI
EIV	The mean of energy value
EV	Energy slope

Corresponding Author: Jian Zhang, Engineering and Technology Institute, Dongguan University of Technology, Dongguan, 523808, China

DurationI is calculated from the annotated manual transcriptions that align the audio documents. SpeakingRate is obtained by phonetic forced alignment by HTK. Next, F0 features and energy features are extracted from audio data by using Praat (Boersma and Weenink, 1996).

The lexical feature set contains eight features: LenI, LenII, LenIII, NEI, NEII, NEIII, TFIDF and Cosine. These features are described as follows:

- **Len I:** The number of words in the sentence
- **Len II and Len III:** The previous/next sentence's LenI value
- tf*idf; tf and idf defined as Eq. 1 and 2
- **Cosine:** Cosine similarity measure between two sentence vectors

$$tf = \frac{n_i}{\sum_k n_k} \quad (1)$$

where, n_i being the number of occurrences of the considered word and the denominator is the number of occurrences of all words in a story or meeting.

$$idf = \log \frac{|D|}{|(d_i \supset t_i)|} \quad (2)$$

where, |D| is the total number of sentences in the considered story or meeting. The denominator is the number of sentences where the word t_i appears.

All lexical features are extracted from the manual transcriptions or ASR transcriptions. For calculating length features, The lecture speech transcriptions are segmented into Chinese words.

Rhetorical-state hidden Markov models: Fung *et al.* (2008) had proposed the supervised learning method, Rhetorical-State Hidden Markov Model (RSHMM) for extracting the rhetorical characteristics from lecture speech.

Each sentence of the given document D is annotated as i^* which approximately maximizes $P(r(s_k)=i|\vec{D})$.

$$i^* = \arg\max_{i=1}^{R} P(r(s_k)=i|\vec{D}) \quad (3)$$

where, r() is a mapping function for the rhetorical unit, there are a total of R rhetorical units in a single document.

EXTRACTIVE SUMMARIZATION USING RHETORICAL STRUCTURE

A common summarization approach-extractive summarization-is adopted for composing a summary by selecting salient sentences or segments from a given speech. In this section, this study briefly describes RSHMM-enhanced SVM for lecture speech extractive summarization (Fung *et al.*, 2008). Then a novel non-generative probabilistic framework-Rhetorical-State SVM- is proposed for further improving summarization process.

RSHMM-enhanced SVM: Based on the probabilistic framework, extractive summarization task is equal to estimating $P(c(s_j)=1|\vec{D})$ of each sentence s_j (Fung *et al.*, 2008) proposed a probabilistic framework-RSHMM-enhanced SVM for summarization process. $P(s(s_j)=1|\vec{D})$ is approximated as following expression:

$$P(c(s_j)=1|\vec{D}) \approx P(c(s_j)=1|\vec{D}, r(s_j)=i^*) \quad (4)$$

where, c() is the salient sentence classification function; i^* can be obtained by Eq. 3. Then sentence s_j whether it is a summary sentence or not using a probability threshold is predicted.

$$P(c(s_j)=1|\vec{D}, r(s_j)=i^*) > \text{threshold} \quad (5)$$

The summarizer is modeled by SVM classifier with Radial Basis Function (RBF) kernel as in Chang and Lin (2001). One SVM classifier is built for each rhetorical unit in the RSHMM network.

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \gamma > 0 \quad (6)$$

Rhetorical-state SVM: For making further use of rhetorical information, a novel probabilistic framework-Rhetorical State SVM is designed to complete summarization task. Using conditional probability theory, $P(c(s_j)=1|\vec{D}, r(s_j)=i^*)$ is deduced to:

$$\frac{P(c(s_j)=1, r(s_j)=i^*|\vec{D})}{P(r(s_j)=i^*|\vec{D})}$$

Considering that the RSHMM network contains R rhetorical units in total, $P(c(s_j)=1|, r(s_j)=i^*|\vec{D})$ is modeled by Rhetorical-State SVM (RSSVM) which contains (R+1) classes: R summary sentence classes and one non-summary sentence class.

One RSSVM is built by (R+1) –class SVM classifier with RBF kernel for the whole RSHMM network. Besides, the $P(r(s_j)=i^*|\vec{D})$ is calculated by Eq. 3.

Finally, those sentences which satisfy the following criterion as summary sentences are predicted:

$$\frac{P(c(s_j)=1, r(s_j)=i^*|\vec{D})}{P(r(s_j)=i^*|\vec{D})} > \text{threshold}$$

EXPERIMENTAL SETUP

The lecture speech corpus are collected containing wave files of 111 presentations recorded from the NCMMSC2005 and NCMMSC2007 conferences. Slides (Microsoft Power Point) and manual transcriptions are also collected. Each presentation lasts about 15 min on average. Each presentation was automatically divided into on average 220 segment units. The ASR system runs in multiple passes and performs unsupervised acoustic model adaptation as well as unsupervised language model adaptation (Chan *et al*., 2007) with 70.3% recognition accuracy.

EXPERIMENT RESULTS

In the summarization experiments, 70 presentations are adopted from the lecture speech corpus. Sixty presentations are used containing 3220 sentences as training data and the remaining 10 presentations of 458 sentences as test data. Importantly, sentence boundaries of all the transcriptions are re-segmented manually. One sentence maybe contains several segment units. Then each presentation contains about 50 sentences. The reference summaries are compiled based on the two assumptions: One assumption is that a good summary should consist of salient sentences from each of the rhetorical units (e.g. title, introduction, background, methodology, experiments and conclusion sections in a conference presentation); the other one is that slides (power point) sentences are good summaries of lecture speech presentations.

The extractive summarization experiments are performed on (R = 3) RSHMM network. Binary SVM classifier without rhetorical information are also built as the baseline system. The summarizer's performance is evaluated by the metric ROUGE (Recall Oriented Understudy for Gisting Evaluation) which can measure overlap units between automatic summaries and reference summaries. ROUGE-L (summary-level Longest Common Subsequence) precision, recall and F-measure are used (Lin, 2004). The results are shown in Table 2.

Table 2: Evaluation by ROUGE-L F-measure of summarization performance on the manual sentence segmentation transcriptions using three-rhetorical-unit RSHMM network

Feature set	Baseline	RSHMM	RSSVM
Le	0.507	0.542	0.561
Ac	0.500	0.513	0.524
Ac+Le	0.516	0.558	0.572

Baseline: The single SVM without rhetorical information, RSHMM: RSHMM-enhanced SVM, Ac: Acoustic Features and Le: Lexical features

This study finds that the Rhetorical-State SVM summarizer consistently outperforms the performance of the base-line system and also outperforms that of RSHMM-enhanced SVM summarizer. In addition, the RSHMM-enhanced summarizer consistently outperforms the performance of the baseline system. In the previous work (Fung *et al*., 2008), the same conclusion is obtained. Furthermore, the best performance is achieved by the Rhetorical-State SVM summarizer based on acoustic and linguistic features: ROUGE-L F-measure of 0.572, 5.6% higher than the best performance produced by the baseline and better than the RSHMM-enhanced SVM summarizer.

Besides, from Table 2, the study also finds that good performance can be achieved by the RSSVM summarizer, only using lexical features extracted from automatic transcriptions with 70.3% ASR accuracy, ROUGE-L F-measure of 0.561. As such, the RSSVM summarizer can decrease the effect of recognition errors on extractive summarization compared with RSHMM-enhanced SVM summarizer.

CONCLUSION

This study has presented a novel probabilistic framework-Rhetorical-State SVM for extractive summarization on lecture speech. RSSVM can automatically decode the underlying rhetorical information in lecture speech and summarize the speech data. In this framework, the RSSVM summarizer produced ROUGE-L F-measure of 57.2% which represents a 5.6% absolute increase in lecture speech summarization performance compared with the baseline system without using rhetorical information. The RSSVM summarizer also outperforms RSHMM-enhanced SVM which is directly built on the rhetorical structure extracted by RSHMM. Besides, only using lexical features extracted from automatic transcriptions, the RSSVM produced good performance: ROUGE-L F-measure of 0.561. This finding suggested that the RSSVM summarizer can reduce the effect of recognition errors on extractive summarization compared with RSHMM-enhanced SVM summarizer.

ACKNOWLEDGMENT

This study is supported by the State Key Program of National Natural Science Foundation of China (U0935003), the Natural Science Foundation of Guangdong Province of China (Grant No. S2012040007560) and the Foundation of Guangdong Educational Committee (Grant No. 2012KJCX0099).

REFERENCES

Boersma, P. 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, pp: 132:182.

Chan, H., J. Zhang, P. Fung and L. Cao, 2007. A mandarin lecture speech transcription system for speech summarization. Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, December 9-13, 2007, Kyoto, pp: 467-471.

Chang, C.C. and C.J. Lin, 2001. LIBSVM: A library for support vector machines. Software, 80: 604-611.

Chen, B., Y.M. Yeh, Y. Huang and Y.T. Chen, 2006. Chinese spoken document summarization using probabilistic latent topical information. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing May 14-19, 2006, Toulouse, pp: 1.

Fung, P., R.H.Y. Chan and J. Zhang, 2008. Rhetorical-state hidden Markov models for extractive speech summarization. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, March 31-April 4, 2008, Las Vegas, NV., pp: 4957-4960.

Hirohata, M., Y. Shinnaka, K. Iwano and S. Furui, 2005. Sentence extraction-based presentation summarization techniques and evaluation metrics. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, March 18-23, 2005, Philadelphia, pp: 1065-1068.

Hirschberg, J., 2002. Communication and prosody: Functional aspects of prosody. Speech Commun., 36: 31-43.

Hori, C., S. Furui, R. Malkin, H. Yu and A. Waibel, 2002. Automatic speech summarization applied to English broadcast news speech. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, May 13-17, 2002, Orlando, FL, USA., pp: 1-9-1-12.

Inoue, A., T. Mikami and Y. Yamashita, 2004. Improvement of speech summarization using prosodic information. Proceedings of the Speech Prosody, March 23-26, 2004, Nara, Japan.

Lin, C., 2004. Rouge: A package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches Out, July 25-26, 2004, Barcelona, Spain, pp: 74-81.

Maskey, S. and J. Hirschberg, 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. Proceedings of the European Conference on Speech Communication and Technology, September 4-8, 2005, Lisbon, Portugal.

Maskey, S.R. and J. Hirschberg, 2003. Automatic summarization of broadcast news using structural features. Proceedings of Eurospeech 2003. http://www1.cs.columbia.edu/~smaskey/papers/eurospeech03.pdf.

Murray, G., S. Renals and J. Carletta, 2005. Extractive summarization of meeting recordings. Proceedings of the 9th European Conference on Speech Communication and Technology, September 4-8, 2005, Lisbon, Portugal.

Zhang, J.J., H.Y. Chan and P. Fung, 2007. Improving lecture speech summarization using rhetorical information. Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, December 9-13, 2007, Kyoto, pp: 195-200.