RHETORICAL-STATE HIDDEN MARKOV MODELS FOR EXTRACTIVE SPEECH SUMMARIZATION

Pascale Fung, Ricky Ho Yin Chan, Justin Jian Zhang

Human Language Technology Center,
Department of Electronic & Computer Engineering,
University of Science & Technology (HKUST),
Clear Water Bay, Hong Kong SAR, China
{pascale,zjustin}@ece.ust.hk, ricky@cs.ust.hk

ABSTRACT

We propose an extractive summarization system with a novel non-generative probabilistic framework for speech summarization. One of the most underutilized features in extractive summarization is rhetorical information – semantically cohesive units that are hidden in spoken documents. We propose Rhetorical-State Hidden Markov Models (RSHMMs) to automatically decode this underlying structure in speech. We show that RSHMMs give a 71.69% ROUGE-L F-measure, a 5.69% absolute increase in lecture speech summarization performance compared to the baseline system without using RSHMM. It equally outperforms the baseline system with additional discourse features, showing that our RSHMM is a more refined improvement on the conventional discourse feature.

Index Terms— spoken document summarization, hidden Markov models, speech features, rhetorical information

1. INTRODUCTION

Spoken document summarization is the recognition, distillation and the presentation of spoken documents in a structural text form, to be presented to the user. The challenge of spoken document summarization, other than automatic speech recognition, lies largely in the lack of easily discernable structures in these documents, in the form of titles, subtitles, sentence and paragraph boundaries, punctuations, fonts and styles to help with the interpretation of the underlying semantic information, which in turn are easily accessible to human readers and search engines alike.

On the other hand, spoken documents make up for their lack of structural information in other features that are present in the speech signal, namely acoustic, phonetic and prosodic information. They represent *how* things are said, whereas the actual words spoken (lexical features) are *what* are said. Existing speech summarization systems, including our previous work, have shown that, for different spoken documents, how things are said is often as important as what things are said.

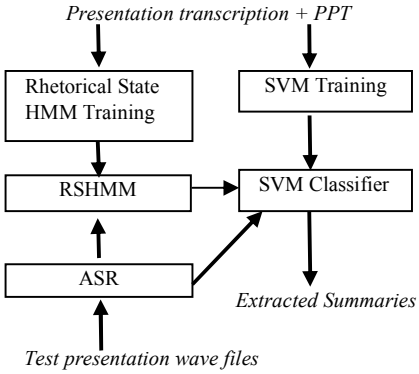

Fig 1. Spoken Document Summarization using Rhetorical-State Hidden Markov Models

Incorporating both acoustic and linguistic features, most spoken document summarization systems employ an extractive approach where salient sentences or segments of the speech are extracted and compiled into a final summary [4,8,9,10,11,13].

Nevertheless, most existing work has failed to make adequate use of one important feature - the rhetorical structure in the spoken documents. Rhetorical structure is the story flow of the document. A document consists of semantically coherent units, known as rhetorical units. A rhetorical unit is a "continuous, uninterrupted span of text" with a single, coherent semantic theme, in a document. In written documents, rhetorical units are often represented as paragraphs or sub-paragraphs.

Our previous work [13] and other researchers have suggested that rhetorical units exist also in spoken documents and efficient modeling of this information is helpful to the summarization task. [9] and [1] used the Hearst method [7] for speech summarization and topic tracking tasks respectively.

Some summarization systems make use of the simplest type of rhetorical information, commonly known as discourse feature, such as sentence or noun position offset from the beginning of the text [1,9,10]. This type of discourse features work well for news reports, but not as well in other genres such as lecture presentations [13].

[4] applied a HMM generative framework to broadcast news speech summarization.

Our proposed work combines the idea of rhetorical structure information and HMM probabilistic framework into summarizing lecture speech presentations. As a result, our HMMs are formulated very differently from the above-mentioned generative model.

This paper is organized as follows: We describe our proposed Rhetorical-state HMMs in Section 2, and the probabilistic extractive summarization framework using RSHMMs in Section 3. Experimental setup and results are in Sections 4 and 5. We then conclude in Section 6.

2. RHETORICAL-STATE HIDDEN MARKOV MODELS

We had previously designed a story flow HMM for text summarization [5,6]. This idea is further developed here for speech summarization.

For a given document **D**, we adopt Hidden Markov Models to represent its rhetorical information. Each of the HMM states corresponds to one rhetorical unit of the document **D**. We define $P(r(\mathbf{S}_j) = i \mid \mathbf{D})$ as the probability for a sentence s_j in the rhetorical unit *i* given the document **D**. We represent each sentence s_j by a feature vector which composes of acoustic and linguistic features. The details of the feature set are described in Table 1.

Table 1: Acoustic and Linguistic feature description for RSHMM

Feature	Description
DurationI	time duration of the sentence
DurationII	the average phoneme duration
LenI	the number of words in the sentence
LenII	the previous sentence' LenI value
LenIII	the next sentence's LenI value
TFIDF	tf*idf; tf is word relative frequency; idf is inverse sentence frequency
Cosine	cosine similarity measure between the sentence vector and the document centric vector

In this paper, we suggest using Rhetorical-State HMMs (RSHMMs) to model the underlying rhetorical structure. Figure 2 shows the concatenation of R RSHMMs to represent a spoken document.

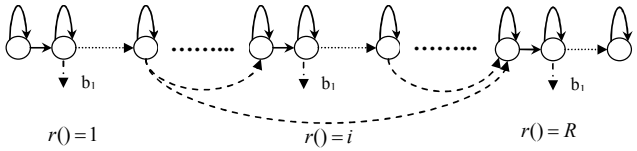

Fig. 2 Spoken document representation with RSHMM

Each RSHMM state contains a probability distribution $b_j()$ for the input feature vector $\mathbf{S}_n$ obtained from the acoustic and linguistic features for the sentence s_n. We used mixtures of multivariate Gaussian distribution as the probability distribution which is represented by the following formula:

$$b_j() = \sum_{m=1}^{M} c_{jm} N(\mathbf{S}_n; \mu_{jm}, \xi_{jm}) \qquad (1)$$

where M is the number of mixture components in the state, c_{jm} is the weight of the m'th component and $N(\mathbf{S}; \mu_{jm}, \xi_{jm})$ is a multivariate Gaussian with mean vector μ and covariance matrix ξ for the acoustic and linguistic features, that is

$$N(\mathbf{S}; \mu_{jm}, \xi_{jm}) = \frac{1}{\sqrt{(2\pi)^n |\xi|}} e^{-\frac{1}{2}(\mathbf{S}-\mu)^T \xi^{-1} (\mathbf{S}-u)} \qquad (2)$$

Considering that the spoken document used in this work are lecture presentations and these presentations consistently follow a rhetorical structure of introduction, content and conclusion, three HMMs (i.e. r_1 , r_2 , and r_3) are built to represent the introduction, content and conclusion section respectively. Each HMM is represented by three states and each of the state contains two Gaussian components. We trained each of the HMMs by performing Viterbi initialization and then followed by Baum-Welch re-estimation using the forward-backward algorithm.

We then place the trained HMMs into a sequential network structure of (r_1, r_2, r_3). We finally use the Viterbi algorithm to find the best rhetorical unit sequence for the document D with N sentence represented by $\{S_1, S_{2,\ldots,} S_N\}$. This is equal to finding the best state sequence $Q^* = \{q_1, q_2, \ldots\ldots, q_N\}$ in the following formula

$$Q^* = \arg\max_Q \{P(S_1, S_2, \ldots S_N \mid r_1, r_2, \ldots, r_R)\}$$
$$Q^* = \arg\max_Q \{a_{q(o)q(1)} \prod_{j=1}^{N} b_{q(j)}(s_j) a_{q(j)q(j+1)}\} \qquad (3)$$

3. EXTRACTIVE SUMMARIZATION

Extractive summarization is a common approach to compose a summary from a document by extracting and concatenating salient sentences or segments from the document. In extractive summarization for spontaneous speech, for a transcribed document D with a sequence of N recognized sentences Sj from the ASR output,

$$D = \{S_1, S_2, \ldots\ldots, S_j, \ldots\ldots, S_N\}, \quad j=1,2, \ldots\ldots, N,$$

we want to find M sentences to be classified as summary sentences by using the salient sentence classification function c().

Based on the probabilistic framework, extractive summarization task is equal to estimating $P(c(\mathbf{S}_j) = 1|\mathbf{D})$ of each sentence s_j. By using conditional probability theorem, it is equal to:

$$P(c(\mathbf{S}_j) = 1 \mid \mathbf{D}) = \frac{P(c(\mathbf{S}_j) = 1, \mathbf{D})}{P(\mathbf{D})} \tag{4}$$

where $c(\mathbf{S}_j)$ is a mapping function for estimating whether the sentence s_j is a summary sentence or not.

Considering rhetorical information of the document **D**, we approximate $P(c(\mathbf{S}_j) = 1|\mathbf{D})$ by the following equation

$$\frac{P(c(\mathbf{S}_j) = 1 \mid r(\mathbf{S}_i) = i, \mathbf{D}) \arg\max_{i}^{R} P(r(\mathbf{S}_j) = i, \mathbf{D})}{P(\mathbf{D})} \tag{5}$$

where $r(\mathbf{S}_j)$ is a mapping function for the rhetorical unit, and we have a total of R rhetorical units in a single document. We then estimate each of the sentence s_j whether it is a summary sentence or not by using a probability threshold. Thus our speech summarization problem becomes finding the sentence s_j which satisfies the following criteria:

$$\frac{P(c(\mathbf{S}_j) = 1 \mid r(\mathbf{S}_i) = i, \mathbf{D}) \arg\max_{i}^{R} P(r(\mathbf{S}_j) = i, \mathbf{D})}{P(\mathbf{D})} > threshold \tag{6}$$

By using conditional probability, equation (6) is equal to

$$\frac{P(c(\mathbf{S}_j) = 1 \mid r(\mathbf{S}_i) = i, \mathbf{D}) \arg\max_{i}^{R} P(r(\mathbf{S}_j) = i \mid \mathrm{D}) \mathrm{P}(\mathbf{D})}{P(\mathbf{D})} > threshold \tag{7}$$

or

$$P(c(\mathbf{S}_j) = 1 \mid r(\mathbf{S}_i) = i, \mathbf{D}) \arg\max_{i}^{R} P(r(\mathbf{S}_j) = i \mid \mathrm{D}) > threshold \tag{8}$$

Extractive summarization is achieved by finding M sentences which give the probabilities higher than the threshold in expression (8). We obtain $r(\mathbf{S}_j)$ for each sentence s_j from the best rhetorical unit sequence which are produced from Viterbi algorithm.

We have previously successfully used sentence vectors for extractive summarization by using support vector machine (SVM) classifier with Radial Basis Function (RBF) kernel. Here, we propose to model $P(c(\mathbf{S}_j)=1 \mid r(\mathbf{S}_j)=i, \mathbf{D})$ by SVM classifier according to the corresponding rhetorical structure units.

Given the best rhetorical unit sequence ($r(\boldsymbol{S}_1)$, $r(\boldsymbol{S}_2)$, ……, $r(\boldsymbol{S}_j)$, ……, $r(\boldsymbol{S}_N)$), then we can find $P(c(\mathbf{S}_j)=1 \mid r(\mathbf{S}_j)=i, \mathbf{D})$ by using the corresponding SVM with the RBF kernel function

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \gamma > 0 \tag{9}$$

4. EXPERIMENTAL SETUP

Our lecture speech corpus contains wave files of 60 presentations recorded from the NCMMSC2005 conference, together with power point files, and manual transcriptions. Each presentation contains about 222 units and lasts approximately 15 minutes. We automatically segment the speech audio into sentence units and produce the transcriptions from our lecture speech transcription system.

The lecture speech transcription system uses tied-state cross-word triphone HMMs. For every shift of 10ms of speech signal, a 25ms window of input speech is represented by 39 dimensions feature vector which compose of 13 MFCC (including C0) and their 1st and 2nd order derivatives. The transcription system runs in multiple passes and performs unsupervised acoustic model adaptation as well as unsupervised language model adaptation [2]. We obtained a 70.3% accuracy for recognizing test data in our experiments.

In our experiments, we use 40 presentations of the lecture speech corpus. We use 85% of the 34 presentations consisting of 6049 sentences as training set and the remaining 6 presentations of 1033 sentences as held-out test set, upon which our summarizer is tested. To compile the reference summaries, we rank sentences from the transcriptions by their cosine similarity to the content of the power points, and then select the top 30% highest ranking sentences. This gives us a compression ratio of 30%.

We train two types of summarizers, with and without smoothing. In the former, a SVM classifier is trained for each state in the corresponding RSHMM. In the latter, A single SVM classifier is trained for each RSHMM. All the HMMs in our experiments are trained by HTK [20] and the SVM models are trained by LIBSVM [4].

5. EXPERIMENTAL RESULTS

We use ROUGE-L (summary-level Longest Common Subsequence) precision and recall [9] as evaluation metrics.

We evaluate the performances of our RSHMM-enhanced extractive summarizer against a baseline summarizer without RSHMM, and against our previous method of using K-

means clustering to find rhetorical boundaries. The results are shown in Table 3.

Table 3: Summarization performance in ROUGE-L F-measure

Features	***Baseline***	***K-means***	***RS HMM***	***RS HMM+S***
Li	.6491	**.6756**	.6190	.7093
Ac	.6195	.5823	.6332	.6717
Ac+Li	**.6600**	.6438	**.6440**	**.7169**

Baseline: Presentation without rhetorical information;
K-means: K-means rhetorical state plus SVM classification
RSHMM: Rhetorical-state HMM plus SVM classification
RSHMM+S: RSHMM with smoothing for SVM classification
Ac: Acoustic; ***Li***: Linguistic

Our RSHMM-enhanced summarizer consistently outperforms the best performance of the other summarizers, as shown in Table 3.

The best performance is achieved by our RSHMM-enhanced summarizer is at ROUGE-L F-measure of 0.7169, 5.69% higher than the best performance produced by the baseline, and better than the summarizer with rhetorical boundaries found by K-means clustering.

In addition, our summarizer also outperforms the baseline summarizer even when the latter uses discourse feature (0.7169 vs. 0.6647 F-measure).

6. CONCLUSIONS

We have presented an extractive summarization system with a novel non-generative probabilistic framework for speech summarization. We proposed Rhetorical-State HMMs to automatically decode rhetorical information in speech. A SVM-based classifier then selects the summary sentences based on their rhetorical states in addition to other acoustic and linguistic features. In this framework, our summarizer produced ROUGE-L F-measure of 0.7169, which represents a 5.69% absolute increase in lecture speech summarization performance compared to the baseline without using RSHMM and is higher than reported in all previous works. We then showed that our RSHMM is even more helpful for summarization task than the conventional discourse feature - 5.22% increase in lecture speech summarization performance.

7. REFERENCES

[1] AKITA, Y., Nemoto, and Y., Kawahara, T., 2007. "PLSA-based topic detection in meetings for adaptation of lexicon and language model." *Proc. Interspeech 2007*, pp.602-605, 2007

[2] H.Y. Chan, J.J. Zhang, P. Fung, and Lu. Cao, "A Mandarin lecture speech transcription system for speech summarization," *To appear in ASRU 2007*, 2007

[3] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," *Software available at http://www. csie. ntu. edu. tw/cjlin/libsvm*, vol. 80, pp. 604–611, 2001.

[4] Y.T. Chen, H.S. Chiu, H.M. Wang, and B. Chen, "A Unified Probabilistic Generative Framework for Extractive Spoken Document Summarization" *Proc. Interspeech 2007*, pp.2805-2808, 2007.

[5] P. Fung and G. Ngai, "One story, one flow: Hidden Markov Story Models for multilingual multidocument summarization," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 3, no. 2, pp. 1–16, 2006.

[6] P. Fung, G. Ngai, and C.S. Cheung, "Combining Optimal Clustering and Hidden Markov Models for Extractive Summarization," *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pp. 21–28, 2003.

[7] M.A. Hearst, "TextTiling: Segmenting Text into Multiparagraph Subtopic Passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[8] M. Hirohata, Y. Shinnaka, K. Iwano, and S. Furui, "Sentence extraction-based presentation summarization techniques and evaluation metrics," *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1, 2005.

[9] C. Hori, S. Furui, R. Malkin, H. Yu, and A.Waibel, "Automatic speech summarization applied to English broadcast news speech," *Proc. ICASSP2002, Orlando, USA*, vol. 1, pp. 9–12, 2002.

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

[11] S. Maskey and J. Hirschberg, "Summarizing Speech Without Text Using Hidden Markov Models," *Proc. NAACL*, 2006.

[12] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.0)," Cambridge University, 2000.

[13] J.J. Zhang, H.Y. Chan, and P. Fung, "Improving lecture speech summarization using rhetorical information" *To appear in ASRU 2007*, 2007