LEARNING DEEP RHETORICAL STRUCTURE FOR EXTRACTIVE SPEECH SUMMARIZATION

Justin Jian Zhang and Pascale Fung

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

ABSTRACT

Extractive summarization of conference and lecture speech is useful for online learning and references. We show for the first time that deep(er) rhetorical parsing of conference speech is possible and helpful to extractive summarization task. This type of rhetorical structures is evident in the corresponding presentation slide structures. We propose using Hidden Markov SVM (HMSVM) to iteratively learn the rhetorical structure of the speeches and summarize them. We show that system based on HMSVM gives a 64.3% ROUGE-L F-measure, a 10.1% absolute increase in lecture speech summarization performance compared with the baseline system without rhetorical information. Our method equally outperforms the baseline with a conventional discourse feature. Our proposed approach is more efficient than and also improves upon a previous method of using shallow rhetorical structure parsing [1].

Index Terms— Rhetorical structure, Lecture speech summarization

1. INTRODUCTION

With the advent of remote learning, distributed collaboration and electronic archiving, there is an increasing need for summarization of presentation speech. This type of speech includes classroom lectures, conference talks, business seminars, as well as political debates and parliamentary speech. Some of the speech are transcribed into text, others might even be accompanied by short abstracts. Nevertheless, for learning and collaboration purposes, transcribed text is too long to read whereas short abstracts do not contain enough information. In our corpus described in Section 4, only about 40% abstract sentences of the conference paper appear in the corresponding transcriptions. Similar phenomenon was found by Teufel and Moens [2]. [3] has shown that extractive summarization is an efficient and effective approach. It has also shown to be more effective than MMR(Maximal Marginal Relevance)-based approach for lecture speech summarization.

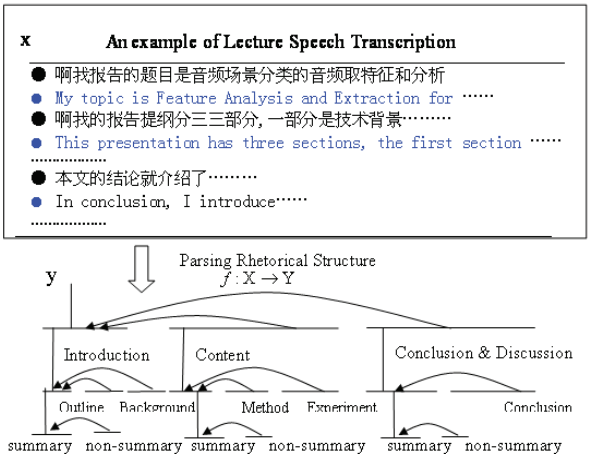

Fig. 1. RST diagram for lecture speech

In recent years, more and more researchers are exploring the hierarchical structure information in a document for better summarization performance [1, 4, 5]. Unlike text documents, the structure of a spoken document is not immediately apparent in terms of titles, subtitles, bullet points, etc. However, [1, 6] showed that structural characteristics of a speech are undeniably rendered by the acoustic and linguistic features of the speech given. As evidenced by the wide spread use of presentation slides with titles, sub-titles, outlines, and bullet points, the hierarchical structure of a document enhances the understanding by the audience. In fact, presentation slides, when available, provide a kind of extractive summarization that is superior in terms of informativeness than short abstracts. Unfortunately, presentation slides are not always made available to the audience or for the archive. In some cases, presentation slides consist of mostly figures and graphs, even videos, with titles and subtitles, but without enough bullet points to summarize the content.

Table 1. Description of Two-dimensional Rhetorical Unit Labels

Attribute	Value
Rhetorical info	(1)Outline; (2)Background; (3)Method; (4)Experiment; (5)Conclusion/Claim
Summary info	(1)Summary; (2)Non-summary

Some summarization systems make use of the simplest type of rhetorical information, commonly known as discourse feature [7, 8, 9]. [1] combines the idea of modeling shallow rhetorical structure using unsupervised learning method K-means and probabilistic SVM framework into summarizing lecture speech presentations. Their experiment result shows that this kind of shallow rhetorical information can help improve summarization process. However, inaccurate rhetorical branch boundaries by K-means tend to be carried over to the summarization step, causing further errors.

Compared to [1, 6] that extract the shallow rhetorical structure which splits a document into "Introduction", "Content", and "Conclusion" parts, we extract deeper hierarchical structure in the document as the rhetorical parse tree, and all the leaf nodes as rhetorical units. Each sentence in the document is first annotated as one kind of rhetorical unit labels described in Section 2, and then chunked into the "Introduction", "Content", and "Conclusion" branches of the rhetorical parse tree as illustrated in Figure 1. We further summarize the document using the deep rhetorical structure. To extract deeper rhetorical structure and overcome the error propagation problem, we design two-dimensional rhetorical unit labels for combining rhetorical structure extraction and summarization into one learning process. One dimension is rhetorical information label; the other is summary information label described in Table 1. We consider this process as multi-class classification process with structured output spaces and build Hidden Markov Support Vector (HMSVM) [10] for the learning process.

This paper is organized as follows: Section 2 describes our motivation, and the rhetorical structure characteristics in lecture speech. Section 3 details how to build HMSVM for learning deep rhetorical structure of the lecture speech and extracting summaries. We then describe the corpus, how to create reference summaries, and the acoustic/prosodic, linguistic and discourse characteristics of lecture speech in Section 4. The experiments and results are presented in Section 5. We then conclude at the end of this paper.

2. RHETORICAL STRUCTURE OF LECTURE

Unlike conversational speech, lectures and presentations are planned. Lecture speakers follow a relatively rigid shallow rhetorical structure at the document level: s/he starts with an overview of the topic to be presented, followed with the actual content with more detailed descriptions, and then concludes at the end. Within each section, there is deeper level of rhetorical structure. For example, the introduction section might start with the motivation, then background. The proposed methodology is followed by an overview of the rest of the presentation. Each of them is in turn a rhetorical unit. These coherent text spans are units of rhetorical structure. Mann and Thompson assert that the structure of every coherent text span can be described by a single rhetorical structure tree, whose top schema application creates a span encompassing the whole document [11]. For lecture speech presentations, we envision the rhetorical structure of lectures and presentations by hierarchical text plan as illustrated in Figure 1. In our work, we use HMSVM for parsing rhetorical structures hidden in speech. We consider this parsing process as multi-class classification with structured output spaces or sequence labeling problem.

Since lecture speeches are mostly based on presentation slides with main gisting points, rather than read from a script, the content and format of the presentation slides is a faithful representation of the document-level rhetorical structure of the lecture speech. To investigate and illustrate this rhetorical structure as represented by acoustic and linguistic characteristics of speech, we use PCA projection of all acoustic/phonetic, linguistic, and discourse characteristics of the lecture speech for visual rendering of the underlying rhetorical structure. We will describe these characteristics in Section 4. PCA reduces the multidimensional feature vectors to two dimensions. We visualize the rhetorical structure of lecture speech, as shown in the upper-left corner of Figure 2. We find that the sentences of the transcriptions are segmented into three sections rather distinctly. This shows that accurate underlying rhetorical structure of the lecture speeches can be obtained by using acoustic characteristics combined with linguistic characteristics. We further annotate the sentences labeled "Introduction sentence" by "Outline sentence" label and "Background sentence" label for representing deeper level rhetorical structure information. Figure 2 shows that these sentences are distinctly segmented into two parts except a few overlaps: the sentences labeled "Outline sentence" in one part and the sentences labeled "Outline sentence" in the other. The similar phenomenon happens in the sentences labeled "Content sentence". This shows that underlying deeper rhetorical structure of the lecture speeches is possibly extracted by using acoustic characteristics combined with linguistic characteristics.

3. LEARNING RHETORICAL STRUCTURE WITH HIDDEN MARKOV SVM

In this section, we will describe how to build Hidden Markov Support Vector Machine (HMSVM) [10] for learning deeper rhetorical structure and extracting summaries from speeches and transcriptions.

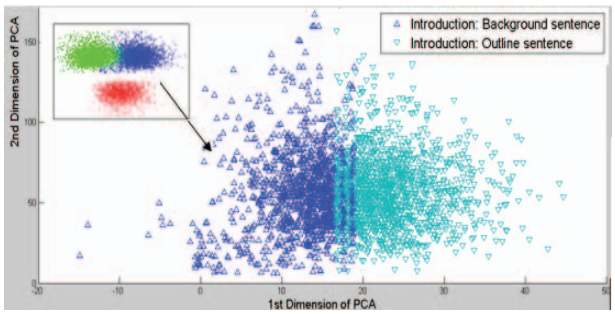

Fig. 2. Visualization of "Outline" part and "Background" part of lecture speech

3.1. Joint Feature Functions

Given the general problem of learning functions $f : \mathcal{X} \rightarrow \mathcal{Y}$ based on a training sample of input-output pairs. As an illustrated example in Figure 1, we consider that the function f maps a given speech or transcription $\mathbf{x}$ to a rhetorical unit sequence $\mathbf{y}$. We intend to find an approach for learning a $discriminant\ function\ F : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{R}$ over input/output pairs from which we produce a prediction by maximizing F over the output variable for a given input $\mathbf{x}$. Equation (1) the general form of our hypotheses f, where $\mathbf{w}$ denotes a parameter vector. We assume F to be linear in some combined feature representation of inputs and outputs $\Psi(\mathbf{x}, \mathbf{y})$ in Equation (2).

$$f(\mathbf{x}; \mathbf{w}) = \underset{\mathbf{y} \in \mathcal{Y}}{\operatorname{argmax}} F(\mathbf{x}, \mathbf{y}; \mathbf{w}) \tag{1}$$

$$F(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle \tag{2}$$

3.2. Hidden Markov SVM

For a transcribed document D, we build an HMSVM for choosing one of the ten kinds of two-dimensional rhetorical unit labels (*Rhetorical-info, Summary-info*) described in Table 1 for labeling all the sentences in D by using optimal function $F(\mathbf{D}, \mathbf{y})$, where $\mathbf{D}$ is a recognized sentence vector sequence $\{\mathbf{s}_1, \ldots, \mathbf{s}_n, \ldots, \mathbf{s}_N\}$. $\mathbf{s}_n$ obtains from the acoustic, linguistic and other characteristics of the sentence s_n. These labels are organized in hierarchical structure as shown in Figure 1. Referring to the corresponding power point slides, we annotate a reference label sequence for each document for building the HMSVM.

For a training example, we generalize the notion of a separation margin by defining its margin with respect to a discriminant function, F, as Equation (3) [10], where the ξ_i are slack variables to implement a soft margin. The linear constraints in (3) are equivalent to the following set of nonlinear constraints: $F(\mathbf{s}_n, \mathbf{y}_n) - \max_{\mathbf{y} \in \mathcal{Y}_n} F(\mathbf{s}_n, \mathbf{y}) \geq 1 - \xi_n$ for $n = 1, ..., N$. Then the solution $\mathbf{w}^*$ of Equation (3) can be written as Equation (4), where $\alpha_n(\mathbf{y})$ is the Lagrange multiplier of the constraint involving example n and labeling $\mathbf{y}$.

$$\begin{aligned} &\min_{\xi \in \mathbb{R}^N, \mathbf{w} \in \mathcal{F}} C \sum_{n=1}^{N} \xi_i + \|\mathbf{w}\|^2 \\ s.t.\ &\langle \mathbf{w}, \Psi(\mathbf{D}, \mathbf{y}_n) \rangle - \langle \mathbf{w}, \Psi(\mathbf{D}, \mathbf{y}) \rangle \geq 1 - \xi_n \\ &for\ all\ n = 1, ..., N\ and\ \mathbf{y} \in \mathcal{Y} \backslash \mathbf{y}_n \end{aligned} \tag{3}$$

$$\mathbf{w}^* = \sum_{N}^{n=1} \sum_{\mathbf{y} \in \mathcal{Y}} \alpha_n(\mathbf{y}) \Psi(\mathbf{D}, \mathbf{y}) \tag{4}$$

We use a Viterbi-like algorithm for decoding the optimal label sequence. When Ψ can be written as a sum over the length of the sequence and decomposed as Equation (5), where γ is the rhetorical unit label set. $l(\mathbf{D})$ is the length of the sequence $\mathbf{D}$. Ψ is composed by mapping functions that depend only on labels at position i and $i+1$, $\mathbf{D}$ as well as i. The score at position $i+1$ only depends on $\mathbf{D}, i$ and labels at position i and $i+1$(Markov property).

$$\Psi(\mathbf{D}, \mathbf{y}) = (\sum_{i=1}^{l(\mathbf{D})} \Psi_{\sigma,\tau}(\upsilon_i, \upsilon_{i+1}, \mathbf{D}, i))_{\sigma,\tau \in \gamma} \tag{5}$$

After labeling all the sentences as rhetorical units, we combine the sentences whose *Summary-info* dimension is annotated as *Summary* into the summarization result.

4. CORPUS AND CHARACTERISTICS

We have collected a lecture speech corpus containing wave files of 111 presentations recorded from different speakers at the NCMMSC2005 and NCMMSC2007 conferences, together with power point slides, manual transcriptions, and their associated audio data. Each presentation lasts about 15 minutes on average. In the current work we use 71 of the 111 presentations which have well-formatted power point slides for our experiments. Each presentation was manually divided into on average 83 sentences.

Based on the finding that inter-annotator agreement is about 80% when instructed to follow the structure and points in the presentation slides i, we generate reference summaries manually correcting those extracted by the Relaxed Dynamic Time Warping (RDTW) between power point sentences and manual transcriptions [12]. The average compression ratio of the reference summaries is 33%.

We represent each sentence by a feature vector which consists of acoustic characteristics and linguistic characteristics. We extract acoustic characteristics: duration of the sentence, average syllable Duration, F0 information characteristics, energy information characteristics and the following linguistic characteristics: length of the sentence counted by word, TFIDF information characteristics. We extract discourse characteristic Poisson Noun described in [1].

5. EXPERIMENTS AND EVALUATION

We perform 6-fold cross validation experiments on manual transcriptions. First, we select 11 documents as our development set for producing the training parameters of HMSVM and divide the remaining 60 documents into six subsets of equal size. We use five subsets to train summarizers by using several supervised methods, as listed in Table 2 and use the remaining subsets for testing. We then evaluate the performance using ROUGE-L F-measure. The average performance of these 6-fold cross validation experiments is shown in Table 2.

Table 2 shows that rhetorical structure can be used for improving performance of the supervised summarizer. We also show that shallow rhetorical structure is more helpful for summarization task than the conventional discourse feature—6.6% absolute increase in summarization performance. Furthermore, we find that by using deep rhetorical structure, our summarizer gives a further 3.5% absolute increase in performance. It shows that deeper rhetorical structure plays a more important role in the summarization process.

Table 2. Av performance: 6-fold cross validation experiments

	nonRS	sRS	dRS
Ac	.490	.520	.54
Ac+Li	.530	.594	.625
Ac+Li+Di	.542	.608	.643

nonRS: extract summaries without rhetorical structure by binary classifier;
sRS: extract shallow rhetorical structure and summaries by HMSVM;
dRS: extract deep rhetorical structure and summaries by HMSVM;
Ac: Acoustic characteristics; Li: Linguistic characteristics;
Di: Discourse characteristics

6. CONCLUSION AND DISCUSSION

In this paper, we have shown that deep rhetorical parsing of conference speech is possible and helpful to extractive summarization. In view of the fact that deep rhetorical structure in speech is inherently hierarchical, we propose a first approach of HMSVM to parse deep rhetorical structure and extract summaries from lecture speech. The performance of HMSVM is superior to that of the baseline summarization system without rhetorical information. In particular, our system with deep rhetorical structure produced ROUGE-L F-measure of 0.643, which represents a 10.1% absolute increase in lecture speech summarization performance compared to the baseline with the conventional discourse feature. That showed that deep rhetorical structure is even more helpful for summarization task than the conventional discourse feature and shallow rhetorical structure for summarization process.

We are interested in applying our model to Automatic Speech Recognition(ASR) transcriptions and other genres of speech, such as meetings, for future work.

7. REFERENCES

[1] J.J. Zhang, H.Y. Chan, and P. Fung, "Improving lecture speech summarization using rhetorical information," *ASRU2007*, pp. 195–200, 2007.

[2] S. Teufel and M. Moens, "Summarizing scientific articles: experiments with relevance and rhetorical status," *Computational Linguistics*, vol. 28, no. 4, pp. 409–445, 2002.

[3] Y. Furui, K. Yamamoto, N. Kitaoka, and S. Nakagawa, "Class Lecture Summarization Taking into Account Consecutiveness of Important Sentences," *Proceedings of Interspeech 2008*, pp. 2438–2441, 2008.

[4] 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.

[5] R. Barzilay and L. Lee, "Catching the drift: Probabilistic content models, with applications to generation and summarization," *Proceedings of HLT-NAACL*, pp. 113–120, 2004.

[6] P. Fung, R. Chan, and J.J Zhang, "Rhetorical-State Hidden Markov Models For Extractive Speech Summarization," *ICASSP2008. Proceedings*, pp. 4957–4960, 2008.

[7] C.H. Nakatani, J. Hirschberg, and B.J. Grosz, "Discourse structure in spoken language: Studies on speech corpora," *AAAI 1995 Spring Symposium Series: Empirical Methods in Discourse Interpretation and Generation*, pp. 106–112, 1995.

[8] S. Maskey and J. Hirschberg, "Automatic summarization of broadcast news using structural features," *Proceedings of Eurospeech 2003*, 2003.

[9] YT Chen et al., "Extractive Chinese Spoken Document Summarization Using Probabilistic Ranking Models," *Proc. ISCSLP*, 2006.

[10] Y. Altun, I. Tsochantaridis, and T. Hofmann, "Hidden Markov Support Vector Machines," in *Machine Learning-International Workshop Then Conference-*, 2003, vol. 20, p. 3.

[11] W.C. Mann and S.A. Thompson, *Rhetorical Structure Theory: A Theory of Text Organization*, University of Southern California, ISI, 1987.

[12] J.J. Zhang and P. Fung, "Active Learning of Extractive Reference Summaries for Lecture Speech Summarization," *ACL-IJCNLP 2009*, p. 23.