Measuring the Structural Importance through Rhetorical Structure Index

Narine Kokhlikyan[†], **Alex Waibel**[† ‡], **Yuqi Zhang**[†], **Joy Ying Zhang**[‡]
[†]Karlsruhe Institute of Technology
Adenauerring 2
76131 Karlsruhe, Germany
[‡] Carnegie Mellon University
NASA Research Park, Bldg. 23
Moffett Field, CA 94035
narine.kokhlikyan@student.kit.edu, waibel@cs.cmu.edu, yuqi.zhang@kit.edu, joy.zhang@sv.cmu.edu

Abstract

In this paper, we propose a novel Rhetorical Structure Index (RSI) to measure the structural importance of a word or a phrase. Unlike TF-IDF and other content-driven measurements, RSI identifies words or phrases that are structural cues in an unstructured document. We show structurally motivated features with high RSI values are more useful than content-driven features for applications such as segmenting unstructured lecture transcripts into meaningful segments. Experiments show that using RSI significantly improves the segmentation accuracy compared to TF-IDF, a traditional content-based feature weighting scheme.

1 Introduction

Online learning, a new trend in distance learning, provides numerous lectures to students all over the world. More than 19,000 colleges offer thousands of free online lectures[1]. Starting from video recordings of lectures which sometimes also come with the presentation material, a set of processes can be applied to extract information from the unstructured data to assist students in browsing, searching and understanding the content of the lecture. These processes include *automatic speech recognition (ASR)* which converts the audio to text, *lecture segmentation* which inserts paragraph boundaries and adds section titles to the lecture transcriptions, *automatic summarization* that generates a short summary from the full lecture, and *lecture translation* that translates the lecture from the original language to the native language of the student.

The transcription of a lecture generated by the ASR system is a sequence of words which does not contain any *structural* information such as paragraph, section boundaries and section titles. Zhang et al. (2007; 2008; 2010) used acoustic and linguistic features for rhetorical structure detection and summarization. They showed that linguistic features such as TF-IDF are the most influential in segmentation and summarization and that knowing the structure of a lecture can significantly improve the performance of lecture summarization. Our experiments with a real-time lecture translation system also show that displaying the rolling translation results of a live lecture with proper paragraphing and inserted section titles makes it easier for students to grasp the key points during a lecture.

In this paper, we apply existing algorithms, namely the Hidden Markov Model (HMM) (Gales and Young, 2007) to unstructured lecture transcription to infer the underlying structure for better lecture segmentation and summarization. HMM has been successfully applied in early works (van Mulbregt et al., 1998; Sherman and Liu, 2008) for text segmentation, event tracking and boundary detection. The focus of this work is to identify cue words and phrases that are good indicators of lecture structure. Intuitively, words and phrases such as "last week we talked about", "this is an outline of my talk", "now I am going to talk about", "in conclusion", and "any questions" should be important features to recognize lecture structure.

[1] http://www.thebestcolleges.org/free-online-classes-and-course-lectures/

These words/phrases, however, may not be so important content-wise. Thus, content-driven metrics such as the TF-IDF score usually do not assign higher weights to these structurally important words/phrases. We propose a novel metric called Rhetorical Structural Index (RSI) to weigh words/phrases based on their structural importance.

2 Rhetorical Structural Index

RSI incorporates both frequency of occurrences and, more importantly, the position distribution of occurrences of a word/phrase. The intuition is that if a term is a structural marker, it usually occurs at a certain *position* in a lecture. Because the term is mainly about the structure rather than the content of a lecture, it can appear with high *frequency* in lectures that are of different topics. For example, "today we" occurs at the beginning of a lecture and "thank you" usually appears towards the end (Figure 1) no matter whether the lecture is about history or computer science.

We define the RSI of a word w as:

$$\text{RSI}(w) = \frac{1}{\lambda \text{Var}(L_w) + (1-\lambda)\text{idf}(w, D)} \quad (1)$$

where L_w is the random variable of "normalized positions" of a word w in a lecture. For each occurrence of w in a particular lecture d, we divide its position by the length of the lecture $|d|$ to estimate its "normalized position". L_w takes a value between [0, 1]. A value close to 0 indicates this word occurs at the beginning and close to 1 means w is close to the end of the lecture. $\text{Var}(L_w)$ is the variance of the normalized position of a word w. A small $\text{Var}(L_w)$ indicates that w always occurs at certain positions of a typical lecture (e.g., "bye") while a large value means w can occur at any position (e.g., function words "of" and "the").

The second part of RSI is the *inverse document frequency* (idf), or effectively the document frequency since RSI is proportional to the 1/idf term. Lectures, such as different research talks, can vary in content but usually have a very similar structure and share some common structural cues. A good structural cue word should be common to many lectures. idf has been widely used in information retrieval research to assign higher weights to words that occur in just a few documents as compared to common words that occur in all documents. Define the idf of a word w given a collection of lectures D as:

$$\text{idf}(w, D) = \log \frac{|D|}{|\{d \in D | w \in d\}|} \quad (2)$$

$|D|$ is the number of all lectures in the collection and $|\{d \in D | w \in d\}|$ is the number of lectures where w appears. A low idf (w, D) value indicates that word w occurs in many documents and thus is more likely to be a common structural cue.

Combining the variance of normalized position and idf by scaling factor λ, we define RSI as in equation 1. We found 0.9 as an optimal value of λ according our experiments over all data sets. A word w with high RSI value is more likely to be structurally important. Similarly, we can calculate the RSI values for phrases (n-grams) such as "I would like to talk about", "I will switch gear to" and "thank you for your attention".

Table 1 shows examples of n-grams and the calculated variance, idf-scores and RSI values from a collection of lectures.

Table 1: Examples of n-grams with high RSI values which are likely to be structural cues.

n-gram	$\text{Var}(L_w)$	idf	RSI
now	0.0004	0.60	1.04
here	0.0004	0.62	1.03
class	0.0001	2.12	0.90
week	0.0001	2.23	0.89
goodbye	0.0001	3.62	0.80
thank you	0.0003	1.53	0.95
talk about	0.0003	1.90	0.92
dealing with	0.0002	2.00	0.91
today we	0.0003	2.51	0.87
see how	0.0009	2.69	0.85
ladies and gentlemen	0.0008	1.35	0.96
last time we	0.0004	2.22	0.89
here we have	0.0005	2.35	0.88
next time we	0.0002	2.51	0.86

3 Incorporating RSI in Lecture Segmentation

Several algorithms have been developed for text segmentation including the Naive Bayes classifier for keyword extraction (Balagopalan et al., 2012), the

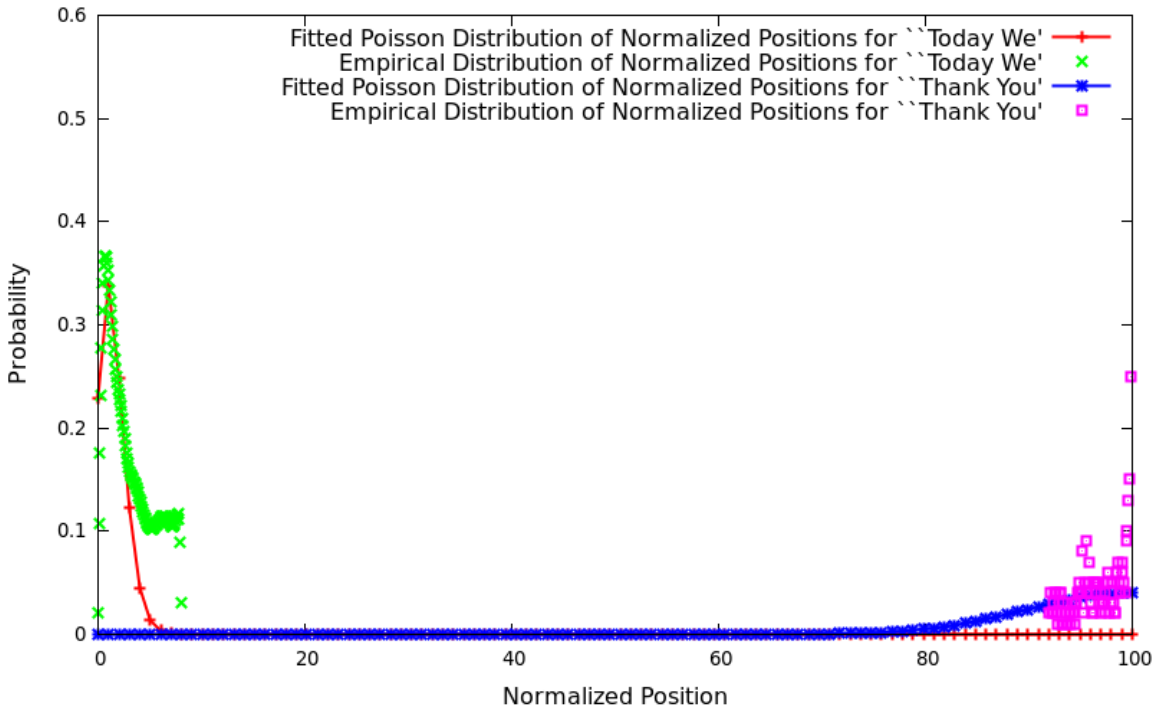

Figure 1: *Fitted Poisson-distribution of normalized positions in lectures for the bigrams "today we" and "thank you". "today we" appears more frequently at the beginning of a lecture, whereas "thank you" more in the concluding part of a lecture. The x-axis is the normalized word position in a lecture and y-axis is the probability of seeing the word at a position.*

Hidden Markov Model (Gales and Young, 2007), the Maximum entropy Markov model (McCallum et al., 2000), the Conditional Random Field (Lafferty et al., 2001) and the Latent Content Analysis (Ponte and Croft, 1997). In this paper, we evaluate the effectiveness of the proposed RSI feature on lecture segmentation using an HMM.

We represent each segment in a lecture as a state in the Markov model and use the EM algorithm to learn HMM parameters from unlabeled lecture data. We use a fully connected HMM with five states. Typical state labels for lecture are: "Introduction", "Background", "Main Topic", "Questions" and "Conclusion" as shown in Figure 2. HMM states emit word tokens. Instead of considering the full vocabulary as the possible emission alphabet, which usually leads to model over-fitting, we only consider terms with high RSI values and high TF-IDF* scores for comparison. For a word w, define its TF-IDF* score as:

$$\text{TF-IDF*}(w) = \max_d \text{TF-IDF}(w, d), \qquad (3)$$

which is the highest TF-IDF score of a word in any document in the collection. Our experiments try to answer the question that "*if HMM is meant to capture the underlying structure of lectures no matter which topic the lecture is about, what kind of features should be emitted from each state to reflect such structural patterns among lectures*?"

The learned HMM model is then applied to unseen lecture data to label each sentence to be "Introduction", "Background", "Main Topic", "Questions" or "Conclusion" and, based on the label, we segment the lecture to different sections for evaluation. Segment boundaries are defined in the positions where sentence labels change.

3.0.1 Bootstrap HMM from K-Means Clustering Segmentation

Initial HMM parameters are bootstrapped using results from K-means clustering where we cluster a sequence of sentences to form a "segment". K corresponds to the number of desired segments of a lecture. Similarities are computed based on the content similarity (using n-gram matches) and the relative

sentence position defined as:

$$Sim(S_i, C_j) = \alpha M(S_i, C_j) + (1 - \alpha) P(S_i, C_j), \quad (4)$$

where S_i is the i-th sentence, C_j is the centroid of the j-th cluster. M(S_i, C_j) is the content similarity between sentence S_i and centroid C_j and P(S_i, C_j) is the position similarity (distance). α is a scaling factor (set to optimal value 0.2 based on all data sets in our experiments).

Content similarity is based on the number of common words between two sentences, or between a sentence and the centroid vector of a cluster. Denote the binary word frequency vector (bag of words) in sentence S_i as $\vec{S_i}$ and similarly $\vec{C_j}$ for cluster centroid C_j, define:

$$M(S_i, C_j) = \frac{\vec{S_i} \cdot \vec{C_j}}{\parallel \vec{S_i} \parallel \parallel \vec{C_j} \parallel}. \quad (5)$$

P(S_i, C_j) measures the position similarity of two sentences. Position similarity is based on the relative position distance between the sentence and the cluster:Define

$$P(S_i, C_j) = \frac{L}{|\text{Pos}(S_i) - \text{Pos}(C_j)| + \epsilon}, \quad (6)$$

where S_i is the i-th sentence, C_j is the j-th cluster. Pos(S_i) is the position of sentence S_i. Pos(C_j) is the average sentence position of all members belong to cluster C_j and L is the total number of sentences in a lecture. ϵ is a small constant to avoid division by zeros.

4 Experiments and Evaluation

We evaluated segmentation on three different data sets: college lectures recorded by Karlsruhe Institute of Technology (KIT), Microsoft research (MSR) lectures[2] and scientific papers[3]. Both college and Microsoft research lectures are manually transcribed. The reason why we do not include experiments on ASR output is that current ASR quality of lecture data is still quite poor. Word-Error-Rates (WER) of ASR output range from 24.37 to 30.80 for KIT lectures. Roughly speaking, every one out of 3 or 4 words is mis-recognized.

[2]http://research.microsoft.com/apps/catalog/

[3]http://aclweb.org/anthology-new/

For evaluation, human annotators annotated a few lectures to create test/reference sets. The test data from KIT is annotated by one human annotator and MSR lectures are annotated by four annotators. The segmentation gold standard is created based on the agreed annotations. Since the number of annotated lectures is small and human annotation is subjective, we also used ACL papers as an additional data set. ACL papers are in a way "lectures in written form" and have titles for section and subsections which can be used to identify the segments and annotate the data set automatically. The statistics of each data set are listed in Table 2.

Table 2: Statistics of three data sets used in the experiments: our own lecture data (KIT), Microsoft research talks (MSR) and conference proceedings from ACL anthology archive. We removed equations, short titles such as "Abstract" and "Conclusion", when extracting text from PDF files from the ACL anthology, which results in a relatively small number of words per paper. Words are simply tokenized without case normalization or stemming, which results in relatively large vocabulary sizes.

Properties	KIT	MSR	ACL
Num.	74	1,182	3,583
Avg. Num. of Sent.	484	655	212
Avg. Num. of Words	10,078	10,225	3,896
Avg. Duration (Min.)	43.57	39.15	-
Vocabulary Size	1.3K	22K	24K

First, we calculate the RSI and TF-IDF* scores for each word in the dataset and choose the top N words as the HMM emission vocabulary. To avoid over-fitting, we choose N that is much smaller than the full vocabulary size of the data set. In our experiments, we set N=300 for KIT, N=5000 for MSR and N=5400 for ACL. The top 5 words with the highest TF-IDF* scores from the MSR data set are: "RFID", "Cherokee", "tree-to-string", "GPU", and "data-triggered", whereas the top 5 words selected by RSI are "today", "work", "question", "now", and "thank", which are more structurally informative.

To estimate the accuracy of the segmentation module, we used Recall, Precision, F-Measure and P_k (Beeferman et al., 1999) as evaluation metric. We used an error window of length 6 to calculate Precision, Recall and F-Measure and a sliding window with a length equal to half of the average seg-

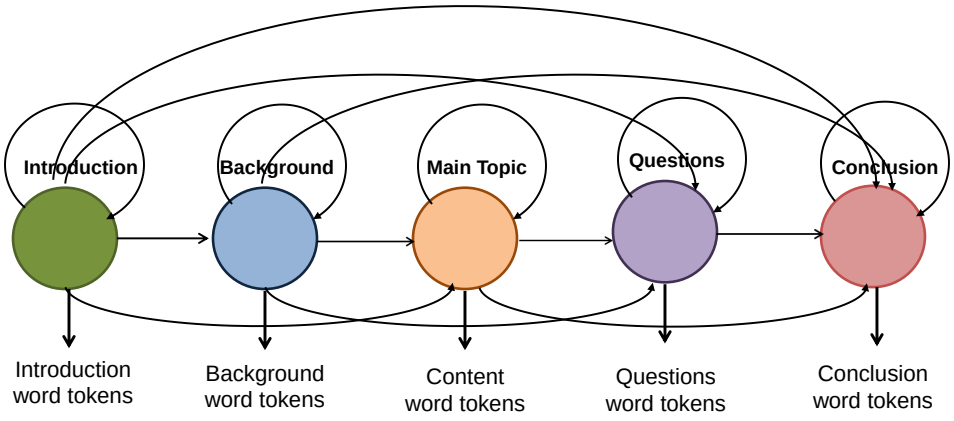

Figure 2: Fully connected 5-state HMM representing *Introduction, Background, Main Topic, Questions, Conclusion* in a typical lecture.

ment length to estimate the P_k score. With error window we mean that hypothesis boundaries do not have to be exactly the same as the reference segment boundaries. Hypothesis boundaries are acceptable if they are close enough to reference boundaries in that window. The P_k score indicates the probability of segmentation inconsistency. Therefore, the lower the P_k score the better the segmentation is.

Table 3: Segmentation results measured by P_k (the smaller the better), Precision, Recall and F-Measure scores (the higher the better) for three data sets comparing HMM using TF-IDF*-filtered word tokens as emission and RSI-filtered words as emissions.

Evaluation Score	KIT	MSR	ACL
P_k			
HMM + TF-IDF*	0.06	0.06	0.05
HMM + RSI	0.01	0.02	0.01
Precision			
HMM + TF-IDF*	32.01	30.47	32.85
HMM + RSI	41.10	41.01	42.70
Recall			
HMM + TF-IDF*	39.32	36.09	38.08
HMM + RSI	47.38	46.39	48.95
F-Measure			
HMM + TF-IDF*	35.29	33.04	35.27
HMM + RSI	44.01	43.53	45.61

The evaluation results on all data sets listed in Table 3 show that according F-Measure and P_k scores, considering words with high RSI values as HMM emission significantly improve over the baseline method of choosing word tokens with high TF-IDF* scores.

5 Conclusions

In this work we propose the Rhetorical Structure Index (RSI), a method to identify structurally important terms in lectures. Experiments show that terms with high RSI values are better candidates than those with high TF-IDF values when used by an HMM-based segmenter as state emissions. In other words, terms with high RSI values are more likely to be structural cues in lectures independent of the lecture topic. In the future we will run experiments on ASR output and incorporate other prosodic features such as pitch, intensity, duration into the RSI to improve this metric for structural analysis of lectures and apply the RSI to other structure discovery applications such as dialogue segmentation.

Acknowledgments

The authors gratefully acknowledge the support by an interACT student exchange scholarship. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287658.

We would like to thank Jan Niehues and Teresa Herrmann for their suggestions and help.

References

A. Balagopalan, L.L. Balasubramanian, V. Balasubramanian, N. Chandrasekharan, and A. Damodar. 2012. Automatic keyphrase extraction and segmentation of video lectures. In *Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on*, pages 1–10.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Mach. Learn.*, 34(1-3):177–210, February.

Mark J. F. Gales and Steve J. Young. 2007. The application of hidden markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289.

Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 591–598, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jay M. Ponte and W. Bruce Croft. 1997. Text segmentation by topic. In *ECDL*, pages 113–125.

Melissa Sherman and Yang Liu. 2008. Using hidden markov models for topic segmentation of meeting transcripts. In *SLT*, pages 185–188.

Paul van Mulbregt, Ira Carp, Lawrence Gillick, Steve Lowe, and Jon Yamron. 1998. Text segmentation and topic tracking on broadcast news via a hidden markov model approach. In *ICSLP*.

Justin Jian Zhang, Ricky Ho Yin Chan, and Pascale Fung. 2007. Improving lecture speech summarization using rhetorical information. In *ASRU*, pages 195–200.

Justin Jian Zhang, Shilei Huang, and Pascale Fung. 2008. Rshmm++ for extractive lecture speech summarization. In *SLT*, pages 161–164.

Justin Jian Zhang, Ricky Ho Yin Chan, and Pascale Fung. 2010. Extractive speech summarization using shallow rhetorical structure modeling. *IEEE Transactions on Audio, Speech & Language Processing*, 18(6):1147–1157.