AUTOMATED LECTURE VIDEO SEGMENTATION:

FACILITATE CONTENT BROWSING AND RETRIEVAL

by

Ming Lin

A Dissertation Submitted to the Faculty of the

COMMITTEE ON BUSINESS ADMINISTRATION

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN MANAGEMENT

In the Graduate College

THE UNIVERSITY OF ARIZONA

2 0 0 6

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation

prepared by Ming Lin

entitled Automated Lecture Video Segmentation: Facilitate Content Browsing And Retrieval

and recommend that it be accepted as fulfilling the dissertation requirement for the

Degree of Doctor of Philosophy

_____ Date: July 25th, 2006
Jay F. Nunamaker, Jr.

_____ Date: July 25th, 2006
J. Leon Zhao

_____ Date: July 25th, 2006
Daniel D. Zeng

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ July 25th, 2006
Dissertation Director: Jay F. Nunamaker, Jr. Date

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

SIGNED: _____MING LIN_____

ACKNOWLEDGEMENTS

I am mostly grateful for my dissertation advisor, Dr. Jay F. Nunamaker, for his guidance, encouragement, and support throughout the five years of PhD study. His advices have been and will always be a fortune for my future career. Thank you to my other major committee members, Dr. Leon Zhao and Dr. Daniel D. Zeng, for their encouragement and support that help made me confident to overcome the difficulties during my study. Thanks to my minor committee members in the Department of Linguistics, Dr. Terry Langendoen and Dr. Mike Hammond for their valuable suggestions and feedback.

Many thanks to the research team at the Center for the Management of Information (CMI) for collaboration and support, as well as the Ford Foundation and the U.S. Air force for sponsoring the projects. Special thanks to: Jinwei Cao, Janna M. Crews, Betty Albert, Chris Diller, Dr. Queen Booker, and Karl Weirs.

I also thank all the faculty members at the Department of Management Information Systems for their advice and support on my PhD study, especially Dr. Mohan Tanniru, Dr. Zhu Zhang, and Dr. Susan A. Brown, and Dr. Alexandra Durcikova.

I would like to thank to my colleagues, fellow PhD students and friends for their support and help through the tough journey, particularly Yilu Zhou, Jason Li, Gang Wang, Yiwen Zhang, Fang Chen, Jie Xu, Harry Wang, and Tiantian Qin.

Finally, I am extremely grateful and greatly indebted to my parents, sisters and brothers, for their unconditional love and support. They are always the energy sources to support me to work hard, and overcome the difficulties during my study and for the rest of my life.

DEDICATION

To my parents

Chuncheng Lin and Anyuan He

TABLE OF CONTENTS

TABLE OF CONTENTS - *CONTINUED*

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

People often have difficulties finding specific information in video because of its linear and unstructured nature. Segmenting long videos into small clips by topics and providing browsing and search functionalities is beneficial for information searching. However, manual segmentation is labor intensive and existing automated segmentation methods are not effective for plenty of amateur made and unedited lecture videos. The objectives of this dissertation are to develop 1) automated segmentation algorithms to extract the topic structure of a lecture video, and 2) retrieval algorithms to identify the relevant video segments for user queries.

Based on an extensive literature review, existing segmentation features and approaches are summarized and research challenges and questions are presented. Manual segmentation studies are conducted to understand the content structure of a lecture video and a set of potential segmentation features and methods are extracted to facilitate the design of automated segmentation approaches. Two static algorithms are developed to segment a lecture video into a list of topics. Features from multimodalities and various knowledge sources (e.g. electronic slides) are used in the segmentation algorithms. A dynamic segmentation method is also developed to retrieve relevant video segments of appropriate sizes based on the questions asked by users. A series of evaluation studies are conducted and results are presented to demonstrate the effectiveness and usefulness of the automated segmentation approaches.

CHAPTER 1

INTRODUCTION

With the advances of Internet and multimedia technologies, a new type of learning system, e-learning systems are becoming popular. These e-learning systems improve the effectiveness of learning and augment distance learning experience by integrating multimedia and Internet content into traditional classroom (Dorai et al 2002, Cao 2005). Among various types of multimedia, video is extremely useful for knowledge sharing and learning because of its great capability of carrying and transmitting "rich" information (Daft and Lengel 1986). Nowadays videotaped lectures are more and more commonly provided in computer-based training systems and they can create a virtual learning environment that simulates the real classroom learning environment. Currently, 84 percent of U.S. colleges offer distance learning programs (Kariya 2003), with a large portion of them integrating lecture videos as a major component of the learning systems. For example, Columbia Video Network (CVN) (http://www.cvn.columbia.edu/) provides six MS degrees and certificate program through its e-learning system and lecture videos are the essential part of the system. University of California at Berkeley has developed on-line learning programs with "Internet classrooms" for various courses in Arts, Science, Business, and engineering (http://learn.berkeley.edu). Similar e-learning systems exist in many other universities such as Stanford University and the University of

Arizona. In addition, videos have been used extensively in corporations as well as for the purposes of medical and military training (Kariya 2003, Smith et al 1999). For instance, General Motors Corp. (Detroit) delivers technical courses to their offices and allows students access the lectures independently as needed. General Motors also videotapes focus groups to help design cars that meet their customers' needs. Although the instructional content may vary in different context, in this dissertation we focus on lecture videos used at universities, considering that the underlying methods could be applied or extended to other types of instructional videos.

Although lecture videos have been used extensively in many e-learning systems, *people often have difficulties finding specific pieces of knowledge in video because of its unstructured and linear nature*. For instance, when students want to review a certain part of a videotaped lecture, they have to look through almost the entire video or play back and forth several times to locate the right spot.

1.1 Lecture Video Segmentation for E-Learning

To address this problem, one idea is to transform and structure the video with other lecture content. For instance, many online courses and e-learning systems use a typical interface that allows students browse different topics of a lecture video structured with other lecture content (e.g. electronic slides). Figure 1.1 shows two examples of such systems: Stanford Online and Agent99 (Lin et al 2003).

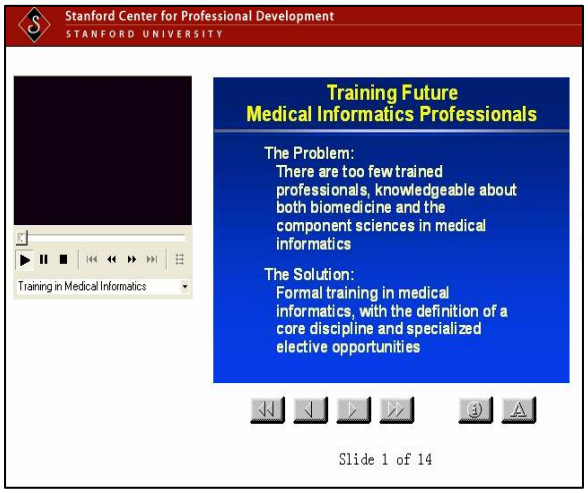


Figure 1.1.a. Stanford Online

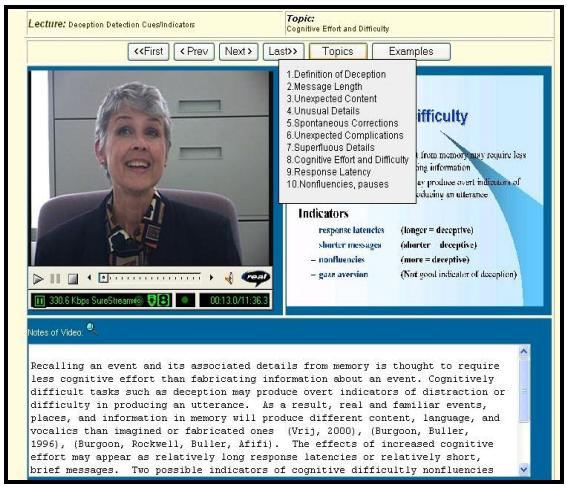


Figure 1.1.b. Agent99 (Lin et al 2003)

In Stanford Online (http://scpd.stanford.edu/scpd/students/onlineClass.htm), a video of an instructor is captured and synchronized with his/her PowerPoint (PPT) slides. Students can move forward or backward to certain segment of the video by choosing the slide associated with that segment (Figure 1.1.a). A similar but improved design was also implemented in two multimedia based learning systems that we developed before: the Learning By Asking (LBA) system (Zhang 2002), and its extension, the Agent99 Trainer system (see Figure 1.1.b) (Lin et al 2003). In both learning systems, each lecture video is manually segmented into short clips, and each clip is synchronized with a PPT slide as well as a text transcript of the speech in the clip. The clips are also indexed based on these text transcripts. Students can select a specific clip in the lecture by browsing a list of topics of the lecture or searching with keywords or questions. Experiment and usability tests have shown that students thought that the list of topics and searching capability

facilitate information seeking. The resulted learning system was as effective as traditional classroom training (Lin et al 2003, Cao 2005).

However, to realize such a structured video lecture and enable browsing and search functionalities, there must be a critical pre-processing step: video segmentation. Without decomposing a lengthy, continuous video into short, discrete, and semantically internal-related video segments, the knowledge structure of the video cannot be extracted, and efficient browsing or searching within the video is impossible. However, performing video segmentation manually is very time consuming because it requires a human to watch the whole video and understand the content before starting to perform the segmentation. Thousands of hours of useful videos are sitting idle in libraries and servers, or cannot be fully exploited because of the huge amount of time efforts required to segment these videos manually. Therefore, there is a strong need to develop automated segmentation algorithms to facilitate the browsing, search, and the full usage of these lecture videos.

1.2 Automated Video Segmentation Algorithms

A variety of video segmentation algorithms have been developed in computer vision, multimedia indexing and retrieval, and Topic Detection and Tracking (TDT) communities (Allan et al 1998, Ngo 2001, Liu and Kender 2004). Depending on the video genres that these approaches focus on, they can be roughly categorized into methods for *films*, *news*, and *lecture videos*. The video genre usually decides the content structure of video, which determinates the effectiveness of certain segmentation method.

A video (from any genre)'s content structure can be viewed as having two levels: *syntactic level* and *event level*. The syntactic level refers to the low level image structure of videos, and the event level refers to the structure based on content instead of the way they are captured or edited (Liu and Kender 2003). For instance, a general content structure of film video is composed of scenes (a distinct story unit), each scene is formed by shots (frame sequence taken at the same site), and each shot is further composed of frames. Thus, most segmentation methods on films concentrate on detecting scene and shot changes in which image cues such as color histogram have been proved to very effective (Wactlar 2000). For news videos, a general structure is composed of stories, and each story may contain one or more shots. Most segmentation approaches on news videos deal with story segmentation; and the special characteristics of news have been largely utilized to facilitate the segmentation (e.g. cue phrase such as "Good morning" is a good indictor of the beginning of a news story).

However, both films and news videos are highly structured and commercially edited, the explicit and implicit rules (e.g. film production rules and formal presentation format of news) of their content construction are great aid on segmentation. Lecture videos instead do not have such a pre-enforced structure as that in film and news. These videos are usually made by non-professionals (e.g. untrained students) and almost have no editing. The syntactic structure of scene and shot in these videos lose much of their meaning because, for instance, a lecture video may have a talking head all through the video.

Furthermore, a formal presentation format such as “Good morning” does not exist for lecture videos. Therefore, methods from *films* and *news* categories are not suitable for lecture videos. Moreover, existing segmentation research on lecture videos either focuses on detecting different presentation formats (e.g. talking head and blackboard) or slide change (Liu and Kender 2003). However, neither the detection of presentation formats nor slide changes capture the meaningful content structure of a lecture video. The research challenge remains on how to extract the event level content structure of a lecture video, more specifically, the topic structure as shown in Figure 1.1.b. Therefore, the first research objective of this dissertation is to develop automated segmentation algorithms to extract the content structure of lecture videos.

Besides segmenting a lecture video into topics, the retrieval of relevant video segment is another research challenge. As indicated by Liu and Kender (2004), “currently there is still no practical system for content-based query and retrieval of lecture videos”. Several universities such as Harvard Business School and University of Arizona have used commercial software from Virage (http://www.virage.com/content/home/) to transcribe, segment and index lecture videos for searching. For instance, the Learning Technology Center at the University of Arizona provides a searchable video library in which students can search videos using metadata (e.g. title, creator, date) and within text transcripts transcribed by automated speech recognition. The system allows users to type in keywords, and the matched words will be highlighted in the transcript. However, there are probably many occurrences of the search keywords and the user still has to browse

through the long transcript. Further, the speech recognition errors may make the transcript hard to read for a human. There is a strong need for a way to retrieve and return relevant video segments to users directly. Our previous system LBA (Zhang 2002) is another attempt of system with search capability. The LBA system allows students to ask a natural language question and the system will return a list of pre-segmented video clips as answers. However, these video clips are pre-segmented manually and are in fixed sizes. The system will fail on answering a user query which requires a smaller or larger segment rather than the pre-divided segment. Therefore, the second research objective of this dissertation is to develop retrieval algorithm to identify the relevant video segments for a specific user query.

1.3 Structure of the Dissertation

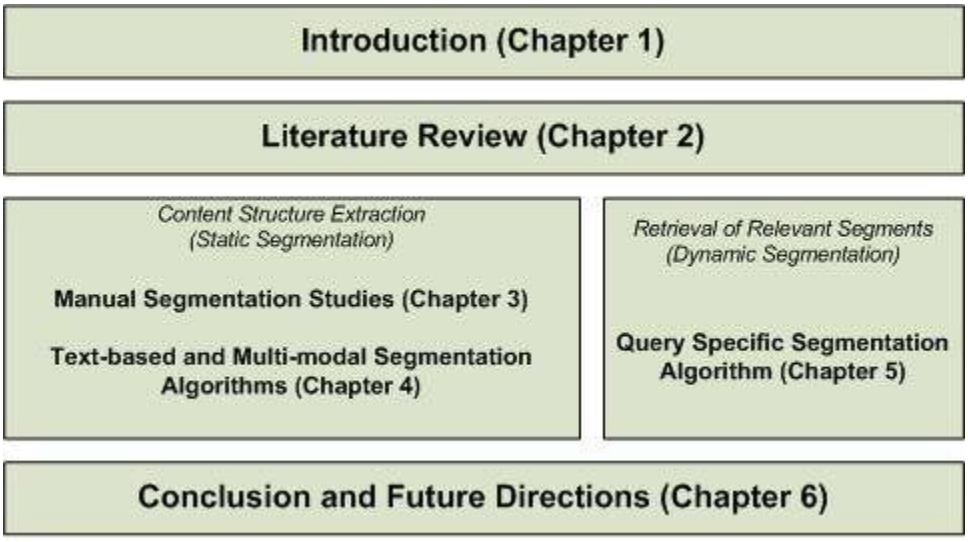


Figure 1.2. Structure of the dissertation

Figure 1.2 depicts the overall structure of the dissertation. Chapter 2 presents a literature review of video segmentation methods on various genres (films, news and lecture videos), followed by a discussion of lecture video characteristics and the details of two research challenges: static segmentation and retrieval of relevant segments.

Chapter 3 and Chapter 4 together present our efforts on addressing the challenge of static segmentation. Chapter 3 discusses two studies on investigating how human experts perform lecture video segmentation manually and collecting rules and heuristics for the design of automated segmentation method. Chapter 4 describes two automated segmentation algorithms. The first algorithm is a text based approach that uses use multiple linguistic based segmentation features (e.g. such as noun phrases and cue phrases) and knowledge sources such as WordNet. The second algorithm combines segmentation features from multiple input sources (speech text transcript, audio and video) and makes use findings from the manual segmentation studies such as two-phase process (initial segmentation + refinement).

Chapter 5 presents a query specific segmentation approach to the challenge of retrieving relevant video segmentation for users. The proposed approach uses a sliding window method to dynamically identify the most relevant segments by computing the similarities between user questions and sliding windows. Extra knowledge source such as electronic slides are used to correct the speech recognition errors in the Automatic Speech

Recognition (ASR) transcripts. Phonetic and partial matching was also utilized to complement the speech recognition errors.

Chapter 6 concludes the dissertation by summarizing the research contributions, limitations, and outlining future research directions.

CHAPTER 2

LITERATURE REVIEW

Video segmentation is the first step towards various video applications such as video browsing, retrieval, and summarization. Simply put, its goal is to divide the video stream into a set of meaningful units as basic elements for indexing. However, what constitute a set of meaningful units has different implications for different video genres because of the variety of underlying content structures. Video segmentation on films is mainly about scene and shot detection, and segmentation on broadcast news focuses on story and shot segmentation (Allan et al 1998, Ngo 2001). Both films and news videos are professionally made and commercially edited. A human enforced syntactic structure is pre-existing from film or news makers. The underlying rules (e.g. film-making rules) have been largely utilized by researchers. On the other hands, lecture videos are usually made by amateur, and lack such a syntactic structure. In other words, the underlying content structure of a lecture video is relatively unclear. Therefore current lecture video segmentation research focuses on detecting different presentation formats (called narrative elements) or slide changing.

This chapter reviews segmentation research in different genres (e.g. films, news, lectures, or text). First, the video segmentation literature on films and news domains is reviewed in Section 2.1. Because the speech text extracted from audio is one major information

source of video, text segmentation research is also presented in Section 2.2. Then Section 2.3 discusses the existing segmentation approaches on lecture videos. Lastly, in Section 2.4, the segmentation methods in various genres emphasizing the special characteristics of lecture videos are summarized, and the challenges of lecture video segmentation are listed.

2.1. Video Segmentation in Film and News Video Genres

Video segmentation on film genre usually focuses on scene and shot detection. Figure 2.1 shows researchers' general view on the content structure of a film type video (Ngo 2001). A video is composed of scenes, which convey distinct story units. Each scene is formed by one or more shots taken place at the same site. Each shot is further composed of frames with smooth and continuous motions. The goal of video segmentation is to restructure the content of videos in a bottom-up manner (Figure 2.1).

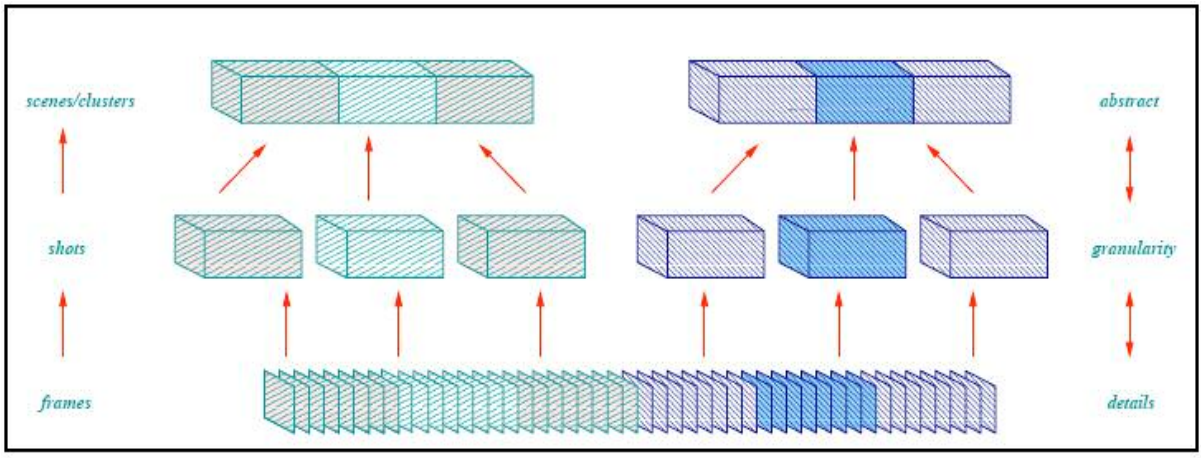


Figure 2.1. General film video content structure (Ngo 2001).

To clarify the concept, we use “video segmentation” in the same sense as “video partition”, not in a sense of segmenting regions and objects in video images as “video segmentation” appears in many video analysis literature.

Research of video segmentation on news videos, also called story segmentation, has been largely motivated by the topic detection and tracking (TDT) initiative (Allan et al 1998). ‘TDT is a body of research and an evaluation paradigm that addresses the event based organization of broadcast news’ (Allan 2000). The goal of a TDT system is to monitor a stream of broadcast news stories, and to find the relationships between these stories based on events that they describe. Story segmentation was a main task of TDT, which is a task of segmenting the stream of data from a source into topically cohesive stories. Many research methods on story segmentation further divide each story into shots (Chaisorn et al 2003, Hsu and Chang 2003). In additional, they classify shots into different categories such as *Intro, Anchor, People, Finance, Weather* and *Commercial*. Figure 2.2 illustrates the general content structure of a news video.

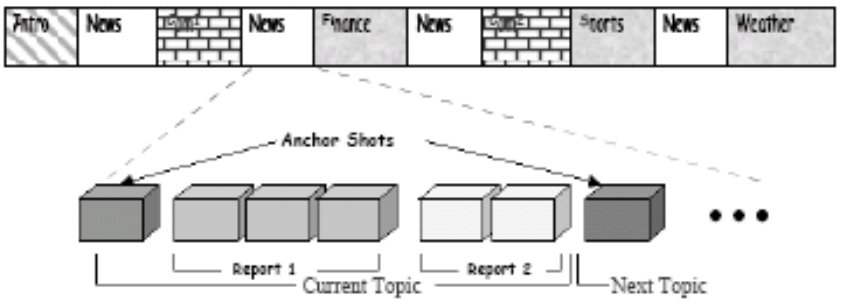


Figure 2.2. General news video content structure (Chaisorn 2003).

The following sections review the segmentation features and boundary identification methods used to reconstruct the above structures in literature. Different research methods are reviewed using a classification based upon the input sources (visual, audio, text, or combination) from which the segmentation features were extracted. Generally speaking, the boundary identification methods can be classified into two big categories: machine learning methods and non-machine learning methods.

2.1.1 Segmentation using Visual Cues

A video (films, news or others) can be partitioned into shots based on visual cues. A shot is an uninterrupted segment of video frame sequence of time, space, and graphic configuration (Ngo et al 2001). The goal of shot cut is to detect the camera breaks or video edits. There are three types of camera breaks: cut, wipe, and dissolve. Most methods on video segmentation utilize color histogram, edge, motion, and statistical hints to identify camera breaks. For example, Wactlar (2000) used color histogram distance computation between successive images to detect scene changes. Zhang and Smoliar (1994) proposed a method for progressive transition detection by combining both motion and statistical analysis. Cut detection algorithms are generally reliable. Color histogram based approaches give superior performance (Gargi et al 2000).

Beside low level feature such as color histogram, segmentation on news video genre also exploits high level object-based visual cues. For instance, Chaisorn et al (2003) extract

the number of faces as well as their sizes to identify anchor shot and shot types such as close-up or long-distance.

Researchers have also explored the relationship between human gestures and topic structure (Quek et al 2000, McNeil et al 2001). Researchers believe that gesture and speech belong to different modalities of human expression and they work together to present the same semantic idea units (Quek et al 2000). Quek et al (2000) used a psycholinguistic device called 'catchment' to integrate different communicative modalities including gesture, speech and gaze components. Their experiment results confirmed the complementary nature of these communicative modalities. Cassell et al (2001) designed a conversational agent that exhibited appropriate posture shifts during dialogue with human users. The authors found that posture shifts occurred more frequently at discourse boundaries than within segments in both monologues and dialogues in their analysis. Most recently, researchers also explored the use of automated recognition and tracking of hand gesture and head pose to assist video editing (Wang et al 2004 and 2005).

Approaches using visual cues from images (e.g. color histogram) are indeed the most common and effective ones for video genres such as films, but they mostly focus on detecting shot and/or scene changes. In general, shot and scene changes detection is not meaningful for lecture videos because a lecture video may have very few shot/scene changes. For instance, in many situations, we may only have a talking instructor all

through the video. Gestures and postures could be potential useful features for video segmentation. But further studies need to be conducted to examine that whether gestures and postures are good indicators for topic shifts in lecture videos.

2.1.2 Segmentation using Audio Cues

Audio track is often a rich source of content information for all kinds of video genres. A large linguistic literature has shown that topic boundaries are indicated prosodically. In other words, major shifts in topic typically show long pauses, an extra high F0 onset, a higher maximum accent peak, and greater range in F0 and intensity. Research has utilized these prosodic features (e.g. pausing, pitch change or rhyme duration) for topic segmentation (Shriberg et al 2000, Tur et al 2001). For instance, Shriberg et al (2000) used a probabilistic model to integrate prosodic and lexical cues for the automatic segmentation of speech into topics. At first a large collection of prosodic features were extracted capturing two major types of speech prosody: duration features and pitch features. A decision tree learning algorithm was used to select salient prosodic features. Then lexical information was captured by statistical language models embedded in a Hidden Markov Model (HMM). The approach is an extension of the segmenter developed by Dragon Systems (Yamron et al 1998), which is based on topic word distributions. Finally, two types of information (lexical and prosodic) were combined and tested in two models: an integrated HMM and a decision tree model which uses HMM posterior as features. Results showed that the prosodic information alone achieved a competitive performance compared to word-based segmentation methods, and combining

prosodic and lexical models achieved the best performance. The results also showed that pause and pitch features were highly informative for segmenting news speech. Besides prosodic features, research also exploited audio cues such as background noise and music. For instance, for *Intro* or *Highlight* shot in a news video, all the narratives are accompanied by background music (Chaisorn et al 2003). However, these types of features are usually domain based and hard to extend to other domains or genres.

2.1.3 Segmentation using Text Cues

Segmentation methods utilizing text input usually make use of transcribed text or closed captions of news videos. With the time stamps that synchronize the video stream and transcribed text (Blei and Moreno 2001), the output of transcribed text segmentation can be mapped back to video segmentation. As discussed before, the majority of research is motivated by the topic detection and tracking (TDT) initiative (Allan et al 1998) with a focus on news and broadcast. Story segmentation, as one main task in TDT, is the task of breaking a broadcast news stream into its constituent news stories.

Unlike written texts, a broadcast news transcript or closed caption does not contain any mark-up indicating where stories begin and end. They lack of topographic cues such as title, paragraph, punctuations, and capitalization. Furthermore, additional speech recognition errors contained in ASR (Automatic Speech Recognition) system output require the segmentation algorithm to be able to filter or handle noisy data.

Many researchers involved in the TDT initiative (Reynar 1998, Beeferman et al 1999) made largely use of the domain cue phrases in news transcripts, which are usually reliable indicators of topic shifts. Phrases such as ‘good morning’, ‘stay with us’, ‘welcome back’ or ‘reporting from PLACE’ normally imply the beginning or ending of a news story. Reynar (1998) identified these phrases and divided them into different categories: ‘Greeting’, ‘Introductory’, ‘Pointer’, ‘Return from commercial’, and ‘Sign-off’ cues (Figure 2.3).

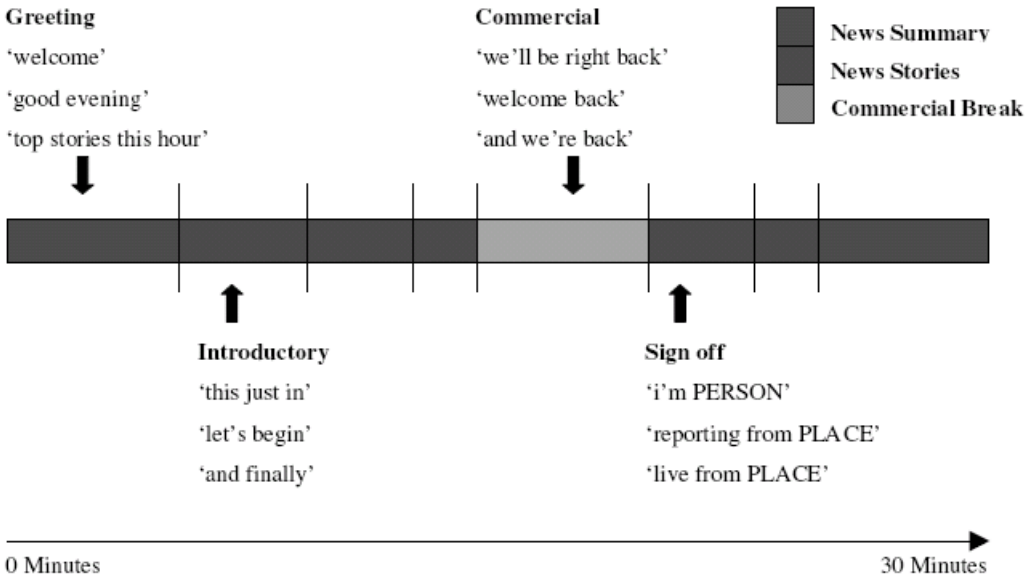


Figure 2.3. A news programme structure and domain cues (Stokes 2004).

One of the main problems with these domain cues, however, is that they are genre and news program specific. For example, European broadcast news does not have “brought to you by PRODUCT NAME” in contrast to American counterparts. Lecture video genre, the focus of this dissertation, even does not have a formal presentation format as a news

program. The varying instructional styles further make it harder to extract specific domain cues.

As illustrated above, many researchers has also exploited multimodal methods to combine visual, audio or text cues together in order to achieve the best performance. For example, Shriberg et al (2000) used probabilistic models (decision and HMM) to integrate prosodic and lexical cues to segment news broadcast into stories. Chaisorn et al (2003) used a combination of features to segment news videos, including visual-based features such as color histogram, object-based features such as face and video text, audio features such as background music, and textual features such as cue-phrases.

2.1.4 Boundaries Identification Methods

The methods used to identify segmentation boundaries generally can be classified to two categories: machine learning methods and non-machine learning methods.

2.1.4.1 Machine Learning Methods

Machine learning methods are most commonly used in segmentation of news video genres. One important machine learning approach that has been successfully used for news story segmentation is Hidden Markov Model (HMM): a method commonly used in speech recognition. In Yamron et al (1997)'s approach, a news story is treated as an instance of a news topic. The approach models a news text stream (with multiple stories) as an unlabeled sequence of news topics in a HMM. Each state in the HMM represents a

topic. A language model is associated with each topic. The probability of any sequence of words can be calculated using the language model. In addition, there are also transition probabilities among the topics. Thus, given a word sequence in a news text, the HMM assigns it a topic with maximum-probability. In the HMM, finding story boundaries is equivalent to finding topic transitions. Unlike Yamron et al (1997) who used only the content based features from the words, Beeferman et al (1997) used the combination of content based features derived from an adaptive language model and lexical features (e.g. domain cue phrases) extracted from discourse structure of the context. He employed a machine learning algorithm called feature induction that incrementally selects the best lexical features and combines them with the language model to form a unified statistical model for the story segmentation. Chaisorn et al (2003) employed decision tree to perform shot detection and classification, and used HMM to identify story boundaries. However, one disadvantage of all these machine learning methods is that they have to be trained and fine-tuned on large set of domain specific data. For a video genre such as lecture videos, the large training data set does not exist yet according to our knowledge.

2.1.4.2 Non-Machine Learning Methods

Most of non-machine learning methods are based on the lexical cohesion theory (Halliday and Hasan 1976). The basic idea is to find topic boundaries by detecting large vocabulary changes. For example, Stokes (2004) developed segmentation system called SeLeCT that uses the number of lexical chains as indictor of vocabulary. A lexical chain is a sequence of related words (e.g. synonyms) in the text, spanning short (adjacent words

or sentences) or long distances (entire text). The SeLeCT system identifies boundaries on locations where a large amount of lexical chains begin or end. SeLeCT uses several types of expanded lexical chains including repetition, synonymy, antonym, generalization/specialization relationships, and part-whole/whole-part relationships (provided by WordNet) (Miller et al 1990). Utiyama and Isahara (2001) proposed a domain-independent statistical model for text segmentation, where no training data was needed since word statistics were estimated from the given text.

2.2. Text Segmentation

Text segmentation literature is reviewed because speech text is one important information source for various video genres, especially for lecture videos. Speech text extracted from audio track conveys most of the information in a lecture video. Furthermore, segmentation features and methods employed in text segmentation are potential features and methods for our lecture video segmentation purpose. As discussing written text segmentation, we concentrate on long documents and unstructured texts. Structured (e.g. data stored in a database) or semi-structure text (e.g. XML) is not our major concerns because the structure information inside them do not exist in the speech texts of lecture videos.

In general the objective of text segmentation is to divide a text document into a distinct set of segments. Segmentation literature has plenty of definitions on what unit of text a segment should represent. These definitions have varied in forms and sizes, from a shift

in speaker focus (a span of speaker utterances) (Passonneau and Litman 1997) to a distinct topical unit like a news story (a set of multiple paragraphs) (Allan et al 1998). Figure 2.4 illustrates the content structure of a text with a top down hierarchy of topics/stories at the top, sub-topics/paragraphs in the middle, and utterances at the bottom. Fine-grained text segmentation, also called discourse segmentation, identifies structure and examines interdependencies between utterances (words, phrases, or clauses). Coarse-grained text segmentation, also called topic segmentation which is the focus of this dissertation, aims to break a text into multi-sentence or multi-paragraph sized text chunks (Figure 2.4).

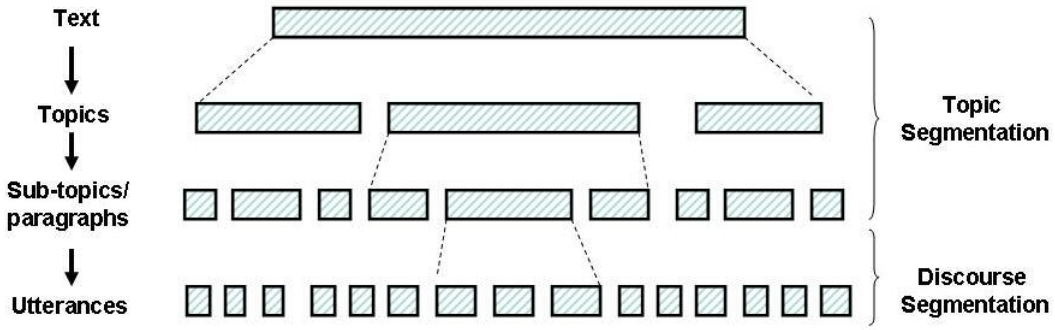


Figure 2.4. General content structure of a text.

2.2.1 Text Segmentation Features and Methods

Most existing work in domain-independent text segmentation has been derived from the lexical cohesion theory suggested by Halliday and Hasan (1976). They proposed that text segments with similar vocabulary are likely to be in one coherent topic segment. Thus, finding topic boundaries could be achieved by detecting topic transitions from cohesion

change. This subsection reviews the literature by showing various segmentation features, different similarity measures used, and various methods of finding boundaries.

Researchers used different segmentation features to detect cohesion. Term repetition is a dominant feature with different variants such as word stem repetition (Youmans 1991, Hearst 1994, Reynar 1994), word n-gram or phrases (Reynar 1998, Kan et al 1998), and word frequency (Reynar 1999, Beeferman et al 1997). The "first uses of words" feature is also used by some researchers (Youmans 1991, Reynar 1999) because a large percentage of first-used words often accompany topic shifts. Finally, cohesion between semantically related words (e.g., synonyms, hyponyms, and collocational words) could also be captured using different knowledge sources such as thesaurus (Morris and Hirst 1991), dictionary (Kozima and Furugori 1993), or large corpus (Ponte and Croft 1997, Kaufmann 1999).

To measure the similarity between different text segments, researchers used vector models (Hearst 1994), graphic methods (Reynar 1994, Choi 2000, Salton et al 1996), and statistical methods (Utiyama 2000). However, only the vector space model is described in this section, other similarity measures are illustrated in the following sections when we present the details of several text segmentation approaches. The vector space model is one of the most popular approaches used by researchers in the Information Retrieval (IR) community. In this model, documents and queries are represented as vectors in *n*-dimensional space. The basic idea is that documents and queries that are similar will be

closer to each other in the vector space than dissimilar documents. The similarity between a query and a document can be calculated using the cosine of the angle between the document vector and the query vector.

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} \tag{2.1}$$

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} (w_{i,j})^2} \times \sqrt{\sum_{i=1}^{t} (w_{i,q})^2}} \tag{2.2}$$

Where $|\vec{d}_j|$ and $|\vec{q}|$ are the norms of the document and query vectors. Both the query and document vectors are weighted. The value of this weight will depend on the weighting scheme used.

The methods of finding topic boundaries in text segmentation include sliding window (Hearst 1994), lexical chains (Morris and Hirst 1991, Kan et al, 1998), dynamic programming (Ponte and Croft 1997, Heinonen 1998), and agglomerative clustering and divisive clustering (Yarri 1997, Choi 2000). Youmans (1991) designed a technique based on the “first uses of word types”, called Vocabulary Management Profile. He pointed out that first-used words frequently followed topic boundaries. Kozima and Furugori (1993) devised a measure called Lexical Cohesion Profile (LCP) based on spreading activation within a semantic network derived from an English dictionary. The segment boundaries can be detected by the valleys (minimum values) of LCP.

2.2.2 Hearst: TexTiling

Hearst (1994) developed a technique called *TextTiling* that automatically divides long expository texts into multi-paragraph segments using the vector space model. Topic boundaries are placed where the similarity between neighboring blocks is low. The algorithm has three steps.

The algorithm first tokenizes a text document and removes stop words. The remaining words are then reduced to their morphological forms. Next, the text is divided into groups called pseudo-sentences of a pre-defined size *w*, which is a parameter of the algorithm. Using pseudo-sentences rather than paragraphs eliminates the difficulties with the vector space model stemming from length variations during similarity measuring. After tokenization, *TextTiling* uses the cosine similarity metric to measure lexical similarity between adjacent blocks of the text, where words are weighted with respect to their frequency within the block. Similarity scores are then calculated for each block gap based on the similarity between a block and its neighboring blocks in the text using the cosine measure (2.2). These scores are plotted against the gap numbers in a similarity graph (Figure 2.5). With the similarity scores, the algorithm then calculates the depth scores. Depth values are computed using the following steps summarized by Stokes (2004):

1. Find the similarity at gap n, i.e. similarity between block n and block n+1.
2. Find the similarity between n and every block to the left of it until the similarity decreases. Record the difference between the similarity at gap n and the highest encountered similarity.

3. Repeat this procedure for block n+1 comparing it to every block on its right.
4. The depth score for this gap is the sum of the two differences calculated in steps 2 and 3.

Simply put, a depth score is basically the sum of the differences between a valley and its immediate left and right peaks.

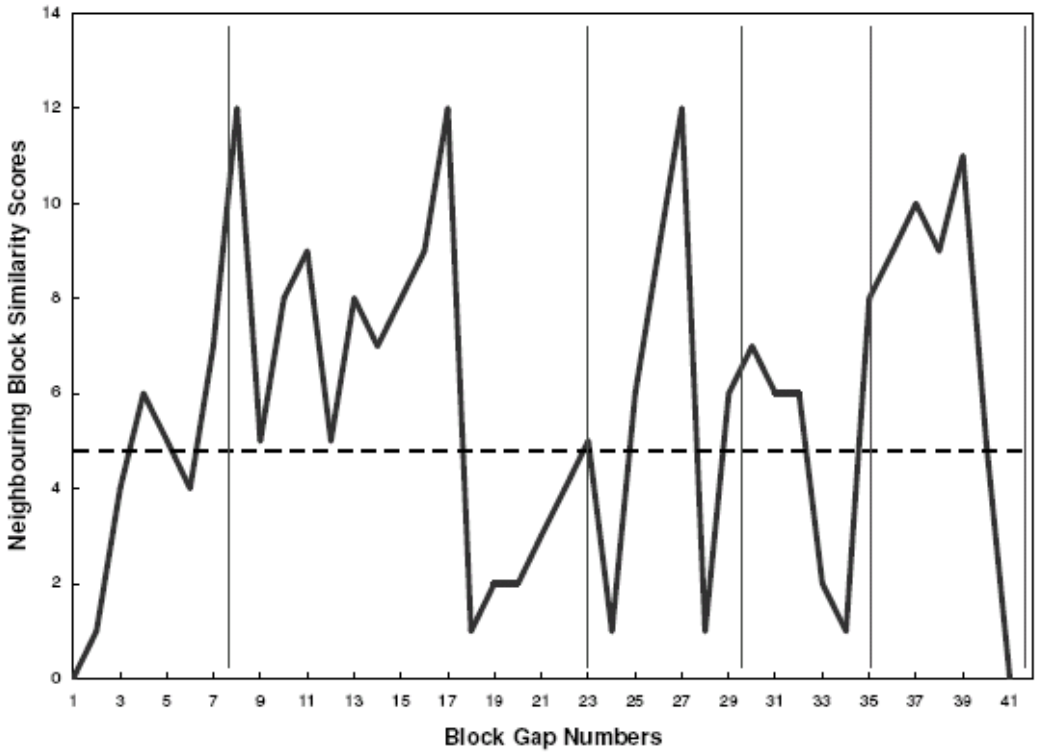


Figure 2.5. Similarity graph determined by *TextTiling*.

Finally, segment boundaries are assigned to gaps with the largest corresponding depth values. These largest values represent areas in the text that exhibit major drops in similarity. The dashed horizontal line in Figure 2.5 represents the cut-off point above

which all depth scores are segment boundaries. This cut-off value is used to decide how many segments to be assigned to a text. Hearst uses a function of the average and standard deviations of the depth scores for the text under analysis. Boundaries are also adjusted to a nearby paragraph breaks. Vertical lines in Figure 2.5 are the boundaries chosen by the *TextTiling* algorithm. It also discards the boundaries that are too close to previously identified boundaries.

2.2.3 Reynar

Reynar (1994) described an optimization algorithm based on word repetition and a graphic technique called dotplotting. To segment a text, the proposed algorithm generates a matrix in which cells (x, y) are set to 1 when word number x and y are the same or have the same root. Cell (x, y) where x = y will always have the value 1 because words are identical to themselves. Figure 2.6 depicts the dotplot of a matrix built by this way.

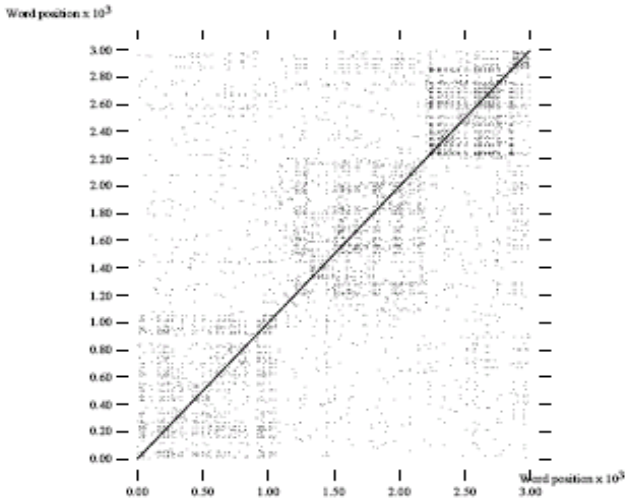


Figure 2.6. An example of the dotplot from (Reynar 1994).

Because the repetition of lexical items occurs more frequently within regions of a text which are about the same topic, the visual appearances of squares along the main diagonal of the plot correspond to regions of the text. Regions are delimited by squares because of the symmetry present in the dotplot. The algorithm identifies topic segments by maximizing the density of the region within the squares along the diagonal or minimizing the density of regions off the diagonal. Another research from Choi (2000) used an approach similar to Reynar's approach (1994). The primary distinction is that inter-sentence similarity is replaced by rank in local context, and boundaries are discovered by divisive clustering.

In (Reynar 1998), Reynar designed two algorithms for topic segmentation. The first is based solely on word frequency, represented by Katz's G model (Katz, 1996). In the word frequency algorithm, a language mode is used to estimate the probability of a sequence of words. The algorithm makes use of *Burstiness*, one linguistic phenomenon that one appearance of a bursty word is a good indicator that additional occurrences are likely. The algorithm uses a language model to determine whether a topic boundary appears between neighboring text blocks. The language model is used to compute the probability of seeing the words in block 2 conditioned on block 1. If the probability of generating the words in block 2 is sufficiently greater when conditioning on the words in block 1 than without conditioning on block 1, the two text blocks are probably about the same topic. Otherwise, the two blocks are most likely about different topics and a putative boundary can be proposed between them.

The second algorithm combines the word frequency with other sources of evidence and incorporates these features into a statistical model built with Adwait's maximum entropy modeling tools (Adwait 1997). The proposed model predicts the probability that a topic boundary is present at a particular location using the following features (Reynar 1998):

- Did the word frequency algorithm suggest a topic boundary?
- Were any domain cues in a cue list present?
- How many word bigrams occurred in both the region before and the region after the proposed topic boundary?
- How many name entities were common to the regions before and after the putative topic boundary?
- How many words in the two regions were synonyms according to WordNet?
- What percentage of words in the region following the putative boundary was used for the first time?
- Were any of the pronouns (from a pronouns list) present in the first 5 words following the potential topic boundary?
- How many words were in the previous segment?

The model is then trained on a subset of HUB-4 1996 Broadcast News Corpus (Reynar 1998).

2.2.4 Ponceleon and Slaney

In (Ponceleon and Srinivasan 2001), an algorithm for the segmentation of video into topically cohesive segments based on ASR transcriptions was proposed. The algorithm is based on discourse structure, boundary conditions between segments, and lexical content. It first extracts word n-grams using frequency counts. Each word n-gram is run as a query against a combined word and phonetic index to obtain occurrences and probability retrieval scores. Then it runs a two-pass segmentation process. In the boundary based first pass, the temporal proximity and the rate of n-gram feature arrival is analyzed in order to compute the initial segmentation. Sufficiently large gaps are identified as initial segmentation boundaries. In other words, changes in the rates of arrival of features are treated as potential topic shifts. In the content-based second pass, changes in content words are detected. The second pass validates the initial segmentation and contiguous segments covering roughly the same topics are merged. This idea of using the rates of feature arrivals is similar to the "first introduce of word types" proposed by Youmans (1991) except that the word types are replaced by n-gram features.

Slaney and Ponceleon (2001) designed a signal processing algorithm to discover the hierarchical structure of a document. Latent semantic indexing is used to describe the semantic content of the signal, and scale space segmentation is used to describe its features at different scales. It is similar to the methods proposed by Choi (2000) and Reynar (1998), in which the task is to search for and identify the square regions of a self-

similar matrix (Section 2.2.3). The algorithm further uses scale-space methods to automatically find the edges of these regions and characterize their strength.

2.3. Lecture Video Segmentation

With the development of internet and multimedia technologies, e-learning systems are becoming more and more popular, and a large portion of them include videos as the major media for content delivery (Abowd 1998, Dorai et al 2001). Detecting and recognizing the content structure, or so called video segmentation, of these lecture videos is the necessary and first step for the browsing and retrieval of video content. However, many existing lecture video based learning systems rely on human labor to perform the segmentation task. For instance, in Classroom 2000 (Abowd 1998), instructors provide presentation slides before class for annotation purposes, and slides are associated with video by annotators.

While automated methods are desired, the majority of automated lecture video segmentation research focuses on either extracting different presentation formats (also called narrative elements) or detecting slides changes. Researchers from these two categories have different views on the content structures of a lecture video as illustrated in Figure 2.6 and 2.7, respectively. The details of two content structures will be illustrated in the following sections. The following terms will be used interchangeably when reviewing literatures: “educational videos”, “instructional videos”, and “lecture videos”.

2.3.1. Narrative Segmentation

Lecture videos usually have various presentation formats: blackboard presentation, handwritten slide, discussion, talking head, electronic slides etc , as shown in Figure 2.7a. The goal of one type of lecture video segmentation research is to detect the change of different presentation formats, also called narrative elements.

Figure 2.7a. Different types of presentation formats in lecture videos (Liu 2004).

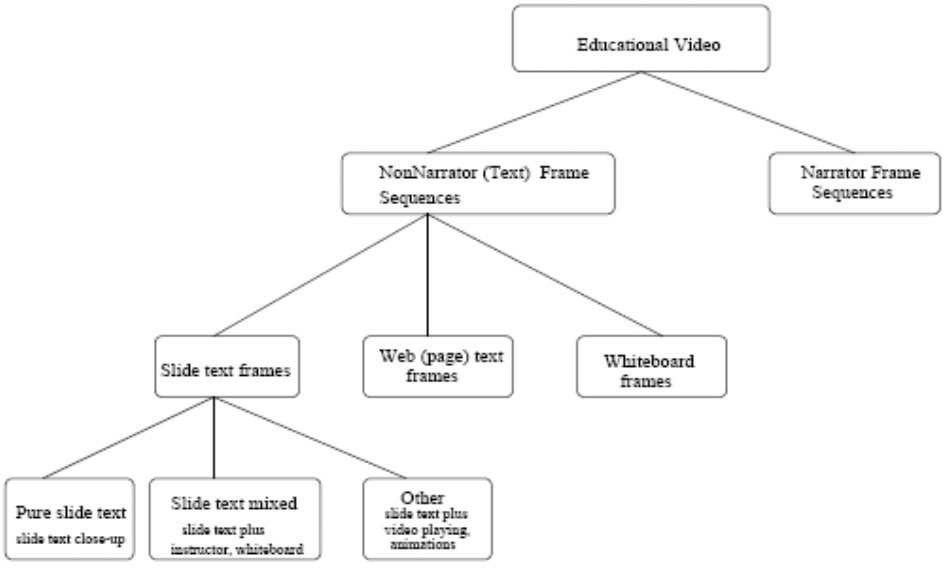


Figure 2.7b. General content structure of a lecture video from a narrative segmentation view (Dorai et al 2003).

Dorai et al (2003) classify sections in lecture videos to narrative elements such as narrator frame sequence, web text frames, and slide text frames. A narrative structure in lecture video (called educational video in this paper) is shown in Figure 2.7b. The top level segmentation involves separating frame sequence to narrator sections (presenter or audiences) and non-narrator (text) frame sequence. The non-narrator sections can be classified into slide text, web text, and white board sections. The slide text sections contains close-up slides, mixed slide shots that may have slides and presenter in the frames together, or slides and white board are both partially seen (Figure 2.7a). A color moments-based feature classification method is used to separate groups of frames. The color moments are feature representations that emphasize the spread characteristics of the spectral distribution of the color values. Color histogram is used to compute color moments and capture the consistent patterns and variations exhibited by non-narrator text-oriented sections and narrator sections. Liu and Kender (2003) used a sort-merge feature selection method to select best features to detect various presentation formats of instructional videos. The sort-merge feature selection method was proposed because of the sparse training data. A hierarchy of small subsets of features was induced by the sort-merge feature selection. A combination of Fastmap and Mahalanobis distance is used for dimensionality reduction and likelihood determination. Video segments are segmented and categorized simultaneously into different formats. Their temporal boundaries are refined progressively using the feature hierarchy.

2.3.2. Slide Matching

Electronic slides (e.g. PowerPoint) are extensively used in traditional teaching, presentation, and in e-learning systems. For videos with electronic slides, one way of structure extraction is to detect the change of slides and relate slide content to video segment. Figure 2.8 shows the view from researchers focusing on slide changes detection or slide matching, usually disregarding the middle level.

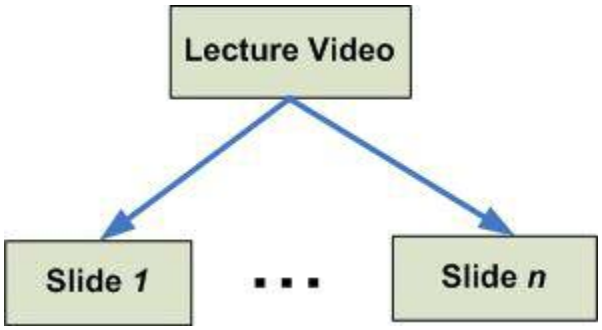


Figure 2.8. General content structure of a lecture video from a slide matching view.

The Cornell Lecture Browser (Mukhopadhyay and Smith 1999) passively captures lecture and structures the lecture videos using the changes in presentation slides. The similarity metric between video frames and slide images is based on Hausdorff distance of two dilated binary images. Ngo et al (2003) proposed an image-based approach to detect the transition of slides. Slides transitions are detected by both the background (figures and design templates) and caption (title and content). A computed background template is used as a mask to detect caption and to compute energy (calculated directly by DC values in a MPEG compress domain) due to background change. A text mask is also generated to compute energy due to caption change. The background and caption

energies are utilized to determine whether a time frame contains slide transitions. However, one weakness of this approach is that position of camera is fixed and it stays stationary.

Liu (2002) presents a slide matching approach based on content differences and dynamic programming. It first detects the "content area" in video frames using a color similarity weighted least square method. The video segments are related to slides by matching the content differences of adjacent video segments to the content differences of all possible slide pairs. After defining the transition probabilities of video segments (based on content differences of binary images), the proposed approach first finds high likelihood matches, and then uses dynamic programming to solve the unmatched remainder.

The topical event detection system (Syeda-Mahmood and Srinivasan 2000) uses multi-modal fusion to detect topics in PowerPoint presentations, called foils in the paper. The slides are used as queries to detect video segments where the slides topics are discussed in the lecture. It uses the image content of slides to detect visual events in displayed slides captured in the video stream; utilizes the textual phrases listed on a slide to detect topical audio events as places where the topical phrases was heard; and finally uses a probabilistic model of event likelihood to combine both visual and audio event detection. The details are described as follows.

1. *Topical video event detection.* To detect slides in video, the approach first cuts the video into shots and each shot is represented by a keyframe. Then the slide matching

is performed between electronic slide images and keyframes. The slide matching or detection has two phases, 1) detecting slide-containing regions in video frames, and 2) recognizing which of a given set of slides appears in a slide-containing region of a video frame. The slide containing regions with frames are detected using the background color of slides. To handle camera pans, movement and surrounding scenes, it makes use of the color and spatial layout geometry of regions on slides using a technique called region hashing. Region hashing is based on the observation that slides displayed on a screen can be modeled as planar regions in space. Then, assuming orthographic projection, the warped, rotated, or scaled slides can be modeled as the affine transforms of the original slide images. Finally, the video event spanned by a slide topic based on image content is taken to be the duration between the scene changes of two consecutive slide matches.

2. *Topical audio event detection.* The topical audio event is defined as the set of contiguous points of time in an audio track where there is a spoken evidence for the phrases listed on a slide. It is detected by a three step algorithm. (1) Individual words on the slides are spotted by merging words and phonetic based retrieval on the speech transcript; (2) Topic phrases are extracted from lines of text on slides. It performs a phrase-based audio retrieval to obtain the places in the video where one or more phrases were heard. The phrase-based retrieval consolidates these matches such that the order of words in the query phrase is preserved in the matching spoken phrases found; (3) The matches to multiple phrases on slides are grouped based on inter-phrasal match distance to identify candidate audio events. It uses a distance threshold

of 20 seconds to group consecutive phrasal matches into time groups using a connected component algorithm. Then the probabilities of individual phrasal match in the group are combined to identify most likely audio events.

3. *Topical slide event detection.* The overall topic slide event is detected by combining the evidences from visual and audio event detection. The combination method exploits the time co-occurrence of individual cue-based event detections and the underlying probabilities of relevance of the time duration to event. The probability that the time duration contains the overall topical event E is given by

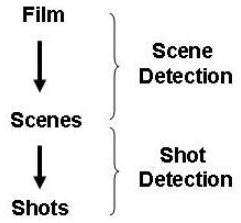
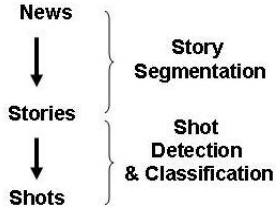
   P(G; E) = P(G; El) + P(G; E2) - P(G; El)* P(G; E2).

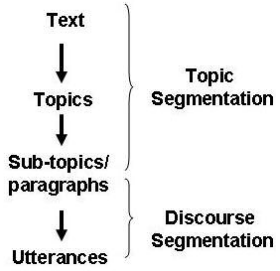
   P(G; El) is the probability of relevance of interval G to event E1 = topic video event.

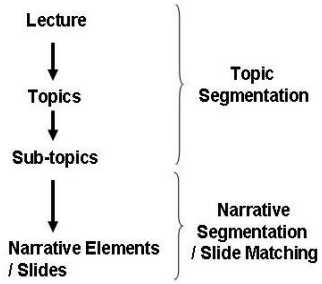
   P(G; E2) is the probability of relevance of interval G to event E2 = topic audio event.

Table 2.1 presents a summarization of all the segmentation literature works we reviewed.

Table 2.1 Summarization of segmentation research

Genre	Content Structure	Segmentation Features	Segmentation Methods	Problems When Applied to Lecture Videos
Films	 *(Figure 2.1)*	Visual cues: color histogram, edge, motion, statistical analysis	• Uses color histogram distance computation between successive images to detect scene changes (Wactlar 2000). • Progressive transition detection by combining both motion and statistical analysis (Zhang and Smoliar 1994).	• Scene/shot changes lose their meaning.
News	 *(Figure 2.2)*	Visual cues: gestures and postures, face detection	• Posture shifts occur more frequently at boundaries (Quek et a 2000, McNeil et al 2001). • Uses the number of faces and sizes to identify anchor shot in news video (Chaisorn et al 2003).	• Gestures and postures are hard to detect and define automatically, need more research. • There is only the instructor's face most of time, but the number and sizes of faces are potential useful for detecting different presentation formats (e.g. instructor narration vs. discussion).
		Audio cues: prosodic cues (e.g. pausing, pitch change or rhyme duration), background noise and music	• Decision tree was used as prosodic model for estimating the posterior probability of a boundary at a given inter-word boundary (Shriberg et al 2000, Tur et al 2001). • Use background music to find intro or highlight shot in a news video (Chaisorn et al 2003).	• Prosodic features are potential useful. • Background noise or music are not applicable because there are no such editing effects in a lecture video.

		Text cues: language model, domain cue phrases (e.g. Good morning)	• Uses decision tree based on multiple features such as word frequency, cue phrase, and pronouns (Reynar 1998). • Uses HMM to estimate the maximum-probability topic sequence based on a language model (Yamron et al 1997).	• Domain cue phrases such as those in news video do not exist in lecture videos because of their amateur and unedited styles. • Machine learning methods are hard to be applied. If applied, the sparse training data problem needs to be addressed at first.
		Combination of features	• Combination of prosodic and lexical features in a probabilistic model (e.g. HMM) (Shriberg et al 2000, Tur et al 2001). • Combination of visual-based features (e.g. color histogram), object-based features (e.g. face and video text), audio features (e.g. background music), and textual features (e.g. cue-phrases).	• The methodology of multi-modal feature fusion is useful for lecture videos. However, features need to be carefully selected and merged so that they can complement each other.
Text	 *(Figure 2.4)*	Term repetition (e.g. word stem repetition, word n-gram or phrases, word frequency), the first use of words	• Similarity measure: vector space model • Boundary identification methods: sliding window (Hearst 1994), lexical chains (Kan et al 1998), dynamic programming (Ponte and Croft 1997, Heinonen 1998), and agglomerative clustering and divisive clustering (Yarri 1997, Choi 2000).	• Compared with written text, speech text has no typographic cues (e.g. headers, paragraphs, sentence mark, and punctuation). • Speech from lecture videos is more spontaneous and causal (compared with speech from news). Speech text does not have formal transition turns, more filler words.

Lecture Videos	 *(Figure 2.7 and 2.8)*	Color histogram, key points, text recognition, spatial layout geometry of slide regions, words and phrase with temporal orders	• *Narrative segmentation* (detection of various presentation formats): uses color moments-based feature classification (Dorai et al 2003), or sort-merge feature selection (Liu and Kender 2003b). • *Slide matching*: similarity of binary images (Mukhopadhyay and Smith 1999), background and caption energies (Ngo et al 2003), color difference of content area and dynamic programming (Liu 2002), topic video & audio event detection (Syeda-Mahmood and Srinivasan 2000). • *Topic Segmentation*: uses chalk pixels to measure the video content on the blackboard, and recognize a topic as a series of instructor's actions that are determined by the temporal & spatial regions of the blackboard (Liu and Kender 2002); topic segmentation by dividing slides (Li and Dong 2006)	• Narrative elements or slide structure do not necessarily match the topic structure of a lecture. • Limitations to specific types of lecture videos (e.g. blackboard or slide videos).

2.4. Challenges That Motivate The Research

Two major challenges remain to be addressed before a practical system for content-based browsing and retrieval of lecture videos can be developed. The first challenge is the effective extraction of the content structure of a lecture video (Liu and Kender 2004). Most existing research in lecture video segmentation only detect narrative elements or slide changes, which is not enough for effective content retrieval. The second challenge is the effective searching of lecture video content. Existing systems are still based on searching using keywords and manually annotations. As indicated in (Liu and Kender 2004), "retrieval of relevant video segments is the most challenging task in fully utilizing instructional videos." The following two sections will further discuss the two challenges.

2.4.1. Content Structure Extraction

It is necessary to recognize and extract the content structure of lecture videos in order to enable effective browsing and retrieval of video content. However, as discussed in section 2.3, existing lecture video segmentation focus on either narrative segmentation or slides matching with very few exceptions. In other words, researchers in the narrative segmentation category view the content structure of a lecture video contains various presentation formats and is a hierarchy with narrator and non-narrator sections (slides, white boards etc) (Figure 2.7a, b). On the other side, researchers in the slide matching category only concentrate on the videos with slides. They view the content structure of a lecture video as a composition of slides, and each slide is a topic.

As researchers (Liu and Kender 2004) pointed out, only detecting narrative elements (different formats) and slide changes or matching is not enough for effective content retrieval. In a lecture, an instructor may explain one topic using a combination of electronic slide, blackboard, and narration. In another lecture, the same instructor may explain several topics using the blackboard. In both cases, a narrative element is not a good representative the actual content structure of the lecture. A semantic structure based on topic, also called teaching topic in Liu and Kender (2004), is more meaningful for content indexing and retrieval. Figure 2.9 depicts my understanding of the content structure of a lecture video.

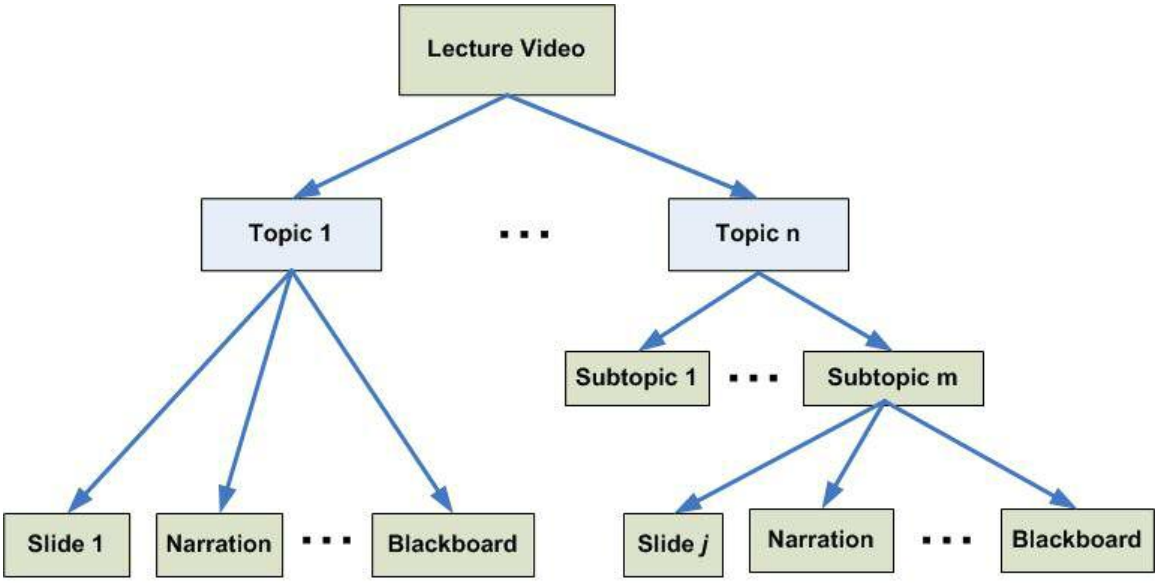


Figure 2.9. General structure of a lecture video

A lecture video contains a series of teaching topics. Each topic can be further divided into subtopics. A topic or subtopic could contain one or several narrative elements such as slide, blackboard, and instructor's narration. In another case, several topics may belong to

one narrative element such as blackboard. Then each of these topics contains only one narrative element: blackboard. There is no need to further divide the topic into different narrative elements. Beside the advantage that a topic is a more meaningful semantic unit for a lecture video, the hierarchical structure in Figure 2.9 has another benefit. The division of a topic/subtopic into different narrative elements is beneficial because the key frames extracted from the narrative elements enclose valuable visual information for user. For example, suppose an instructor use both a slide and blackboard to teach one topic: how does a search engine work? Both the slide image and a key frame of the blackboard complement each other to express the same topic unit. Showing both of them will benefit the end users more.

Therefore, one research challenge of lecture video segmentation is to reconstruct the content structure of lecture video as illustrated in Figure 2.8. While previous narrative segmentation methods (Dorai et al 2003, Liu and Kender 2003) can be applied directly to segment a topic/subtopic into narrative elements, *segmenting a lecture video into topics and subtopics remains a research challenge*. The following two subsections discuss the previous research on the topic segmentation of lecture videos and the research challenges introduced by the special characteristics of lecture videos.

2.4.1.1. Related Work

Liu and Kender (2002) proposed a method to recognize teaching topics in blackboard by extracting spatial-temporal grouping of teaching content based on the recognition of the

significant actions of instructors and the spatial and temporal coherence of blackboard content. The method first segments frames into board regions, instructor occlusion regions, and irrelevant regions. The number of chalk pixels is used a heuristics to measure the video content. The method uses the temporal and spatial regions of the blackboard to recognize a topic as a series of instructor's actions represented as *SWET*. *S* represent the starting of a topic by starting writing on the board; *W* represent instructor writing more lines; *E* represent the action of instructor add more symbols or underline; *T* represent the action of instructor talking. However, this method has two weaknesses:

- One weakness of is that it only targets on blackboard videos. It is hard to generalize to other types of lecture videos such as videos with both slides and blackboard, which are the most common in lecture videos.
- The second weakness is that the instructor actions are limited to *SWET*. Many instructors may not follow the same behavior pattern. For instance, one instructor may erase one line on the blackboard before adding more lines, or starts to write on another side of the board.

Li and Dong (2006) proposed a method to segment presentation video hierarchically. The method employs visual information (local color histogram difference) on the slide-level segmentation. Topic level segmentation is a segmentation of slides text based on topical words introduction. Beside the limitation of slides video and generalization problem, the slide level segmentation requires full size slide images on all the frames in the video in order to make the color histogram differences method work. Thus, a general topic

segmentation method without limitation to specific types (e.g. slide or blackboard videos) is desired.

2.4.1.2. Lecture Video Characteristics

The special characteristics of lecture videos are summarized as follows. Refer to Table 2.1 for details of segmentation research on various genres and their weakness when applied to lecture videos.

- *Non-professional Made and Unedited*: the first non-professional means that the camera man is not trained or well-trained who do poor camerawork. Further more, these videos usually do not go through any editing work. Non-professional also means that the speakers in the lecture or presentation, usually the instructors, are not professionally trained neither, unlike the speakers in news broadcast. Further, there are no formal presentation format requirements for instructions in a classroom. Non-professional made and no editing are the main reasons that make the segmentation of lecture video difficult, and introduce the characteristics in the following bullet points. For example, poor camera work and no post-editing imply that there is no underlying syntactic structure. Non-professional speakers and non-professional format mean that the speech is more spontaneous.
- *No Syntactic Structure*: Lecture videos usually lack scene and shot changes. Scene and shot changes are not meaningful and do not match the actual semantic (topic) structure of lecture videos. Further, there is no notion of a story as in news video segmentation and the associated well-defined segment of expected relatively short

duration. The duration and number of cohesively topical segments can vary between different lectures and instructors. Typically even manual segmentation for such video content can be very subjective and inconsistent (Section 3).

- *Heterogeneous*: Unlike the corpus of broadcast news video (e.g. these used in TDT evaluations), lecture or instructional videos can be quite heterogeneous. For instance, in one lecture, an instructor uses PowerPoint slides alone and follows the slides strictly. While in another lecture, s/he may use blackboard and slides together, and jump back and forth between slides. Furthermore, different instructors tend to have quite various instructional styles. These indicate that segmentation features and methods designed for one instructor or one type of lecture video may not work in another instructor or type of video. For example, previous algorithms designed for slide change detection cannot be applied to blackboard videos.
- *Spontaneous Speech*: The speech inside a lecture video tends to be more causal and spontaneous than that in a news video. Apart from poor planning at the sentence level including more filler words (“okay”, “well”) and non-lexical pauses (“uh” or “um”), lecture speech often exhibits poor planning and structure on higher levels as well, for example, with tangential topic digressing from current primary topic (Glass et al,). Silence and non-speech events tend to be longer between stories. There is no such a structure that the speaker has to make the transitions as in a news talk. Finally, because of the spontaneous characteristic of the lecture speech, background noise, professional versus amateur speaker etc, and the accuracy of the speech recognition

transcripts are likely lower than those in news transcripts, reported as from 35-60% word error rate (Ponceleon and Srinivasan 2001).

- *Extra Knowledge Sources*: Electronic slides and class websites have been extensively used and become more and more popular recently. Electronic version of the textbook is also available in some cases. All the above additional materials provide extra knowledge sources to facilitate the video segmentation and boost the segmentation performance. For instance, Yamamoto (2003) used the textbook associated with the lecture to improve the topic segmentation.

With all the special characteristics of lecture videos, segmentation features and methods working on films, news videos, and text, are not directly applicable for lecture videos. There is a strong need to exploit the salient segmentation features and reliable methods. Because manual segmentation provides the most accurate results and is usually used as a benchmark for segmentation evaluation (Allan et al 1998), I conducted several studies of manual segmentation (Chapter 3). Chapter 4 discusses two algorithms I proposed to address the topic segmentation problem: the first algorithm is text based, and the second one fuses features from multi-modalities.

2.4.2. Retrieval of Relevant Video Segments

Extracting the content structure of lecture videos by topic segmentation is a great aid in browsing and retrieval of video content. However, static topic segmentation in a pre-processing manner only solves the content retrieval problem partially. The most effective

way of retrieving lecture video content is still by searching. Returning a pre-segmented video chunk as the answer for a user query has several disadvantages. Users have various information requirements which probably require returning different segment sizes. For instance, returning a 10-minute pre-segmented video clip may be too much for a user who wants to know the answer of a question such as "*how many calories are there in a big Mac*?" A one or a few sentences shorter video segment is probably a better answer. Therefore, *a retrieval system that can return segments with dynamic size based on user queries is highly desired*.

As indicated by Liu and Kender (2004), "currently there is still no practical system for content-based query and retrieval of lecture videos." There is very few systems designed specifically for lecture videos to our knowledge. (Fujii et al 2003) is one exception. Fujii et al (2003) propose a cross media retrieval system for lecture TV program using text queries. The text queries contain keywords from users or can be formulated automatically by retrieving text passages from the textbook. Lecture videos are pre-segmented into passages ranging from 1 to 5 sentences. Top ranked passages are returned to users using the same method as that for text retrieval. However, the method highly relies on electronic textbooks, which are commonly available for many lecture videos. Further, the return passages still have fixed size.

Therefore, retrieval of relevant video segment with appropriate size based on user queries (keywords or questions) remains a research challenge for lecture videos. Chapter 6

discusses my approach to address the content retrieval problem. Relevant literatures such as video question answering and passage retrieval are also reviewed.

CHAPTER 3

MANUAL SEGMENTATION STUDIES

Although segmenting lecture videos into topics is beneficial, the actual segmentation process itself is not an easy task. While manual segmentation provides the most accurate results and is usually used as a benchmark for segmentation evaluation (Allan et al 1998), it is very time consuming. Automated video segmentation can save labor time, but its accuracy is still far from optimal. As discussed in chapter 2, existing segmentation methods on genres such as films, news video or text are not suitable for lecture videos. Segmenting a lecture video into topics automatically remains a research challenge. Furthermore, most existing literature concentrates on specific types of lecture videos such as slide videos (Ngo et al 2003) or blackboard video (Liu and Kender 2002). There is no research exploring more generalizable segmentation features and methods.

The objective of our research study in this dissertation, therefore, is to design an automated segmentation method suitable for lecture videos with high accuracy. To achieve accuracy rates as high as manual segmentation, it is critical and beneficial to study how humans, especially experts, perform the video segmentation manually. The rules and heuristics that humans use in their segmentation process could be used as foundations for the design of the automated method. This chapter investigates how humans perform manual segmentation and collecting rules for the design of an automated

segmentation method for lecture videos. The rest of this chapter is organized as follows. Section 3.1 discusses an exploratory and pilot study focusing on understanding the problem and challenges. Section 3.2 presents a formal study with a goal to find potential segmentation features and methods. Section 3.3 summarizes the findings from two studies, and the potential segmentation features associating with feature extraction methods from the literature.

3.1. An Exploratory Study of Manual Segmentation

My first manual segmentation study is exploratory and concentrates on understanding the segmentation phenomena. The objectives of study are to understand the human segmentation process, lecture video characteristics, and factors affecting segmentation. I am more interested in whether there are reliable segmentation features across various classes and instructors, and leave the identification of potential features to the next study. More specifically, I am interested in the following research questions:

1. The topic structure of a lecture video
    - Do different people have different perspectives on the topic structure of a lecture video, and why?
    - What are the natural structures of a lecture video, linear or hierarchical?
2. Are there any reliable segmentation features for all type of lecture videos, and what are they?
    - What is a topic transition (e.g. a time point or a short time)?

3. What are the factors (video taping methods, external knowledge sources such as electronic slides and speech transcripts) affecting the video segmentation, and how?

3.1.1 Study Design

In order to answer the above questions, five videos were randomly selected from three MIS courses and assigned to four experts. The four expert participants were asked to segment the videos assigned to them manually, identify the topic transition time points, and most salient features s/he used in each individual boundary. They were also required to record several statistics (e.g. time cost used for the segmentation) and fill out a questionnaire. The questionnaire includes mainly three open-ended questions: "what is their general segmentation process", "what are their perspectives on the topic structures of the lecture videos", and "whether factors such as video taping (e.g. camera settings and professional or unprofessional camera guy), slide and speech transcript affect their segmentation". Automatic Speech Recognized (ASR) transcripts and slides are also provided (if available) to assist the segmentation and test the effects of external knowledge sources.

In order to answer the question of "are there any reliable segmentation features for all type of lecture videos?" I tested the following measures in their specific settings:

- *Segmentation consistency between different people*: use a setting of "same class videos/different participants"

- *Common segmentation features*: use settings of "different classes/same course" and "different classes/different courses"
- *Segmentation consistency between instructors*: use a setting of "different classes/different instructors"

To measure the above measures in their appropriate setting, 1 or 2 class videos were selected from each of the three courses, which are taught by different instructors. Further, the five videos are carefully chosen and assigned to four participants in such a way to ensure that cross comparison is possible (Table 3.1). For instance, in Table 3.1 the same class video (e.g. L1) from the same course (MIS 1) was assigned to two different participants (1 and 2) to check segmentation consistency between different people.

Table 3.1. Exploratory study video segmentation assignments

	Assigned Video 1	**Assigned Video 2**
Participant 1	MIS 1 (L1)	MIS 3 (L4)
Participant 2	MIS 1 (L1)	MIS 1 (L3)
Participant 3	MIS 2 (L2)	MIS 3 (L5)
Participant 4	MIS 1 (L3)	MIS 2 (L2)

3.1.2 Results and Findings

After collecting the segmentation results and questionnaires, results and answers were analyzed and discussed by the four expert participants. The following is a summarization of results and findings to the question of manual segmentation process and the research questions proposed at the beginning of section 3.1.

3.1.2.1. Manual segmentation process

As the manual segmentation processes taken by the four experts vary to some extent, they share certain common features. All of them used a two-step segmentation method when manually segmenting the class videos: 1) watching the video straight through before conducting manual segmentation to know roughly where the breaks are; 2) playing back the video on one screen while making segmentation observations on a second monitor, and at the same time, closely observing and refining the time breaks and topic break outs.

3.1.2.2. Research Questions Answering

The results and findings of the research questions are listed as follows.

1. The topic structure of a lecture video (including perspectives on topic structure and linear vs. hierarchical)

Results and findings: 1) All experts agreed that what constitutes a topic may not be necessarily the same for everybody. Most of time the differences and disagreements are caused by granularity, or in other words, whether to have a coarser division of topics or a finer division of subtopics. 2) The topic structure of a lecture video is viewed as hierarchical by one person whereas may be viewed as linear by another person. Even for the same person, he may be segmenting one lecture in a linear manner but another lecture hierarchically. The participants also report difficulties and cognitive efforts on segmenting a lecture in a hierarchical fashion. It is very hard to recognize the miniscule difference in the various subtopics. Further, the subtopics may not be very coherent when, for example, one teacher may talk about one topic when interrupted by a student

who asked another topic. This interruption makes the topics transition very difficult to catch. In summary, human intends and finds it easier to segment the lectures in a linear fashion rather than in a hierarchical way.

2. Are there any reliable segmentation features for all type of lecture videos, and what are they (topic transitions and consistencies between people, classes, courses and instructors)?

Results and findings: 1) a topic transition could be a time point, but also a short time period. For instance, in one lecture the instructor uses a few sentences to explain the transition: what has been discussed, and what will be discussed later. The participants also reported that in one course, the instructor has a distinct style of transition by some typical sentence structures preceding the introduction of the next topic such as “so, alright, let’s…, now”. While in another lecture, the transition is typically represented by a long pause and changes of the slides corresponding to the topics. 2) The results show that there are no reliable segmentation features across courses and instructor; however there may be reliable features for the same instructor. For instance, as discussed in what are the transitions above, in one course MIS 1, the instructor typically had a topic transition interspersed with some cue phrases like “so, alright, let’s.. , now.” Whereas in course MIS 3 with another instructor, there were no cue phrases like what are in MIS 1. The only salient segmentation cues are long pause (audio cues) and changing slides (video cues).

3. What are the factors (video taping methods, external knowledge sources such as electronic slides and speech transcripts) affecting the video segmentation, and how?

Results and findings: 1) The professional level of camera person did affect the segmentation in a way that unprofessional video taping will make the segmentation harder. For instance, in some videos of MIS 1, the untrained camera person fails to move to the slide on time when the instructor is highlighting the slide content or switching to next slide. The unprofessional video taping makes the video hard to understand and segmentation more difficult when it is supposed to be easy. For example, a slide change in the video is a good indicator for topic change, however, the slide change is not captured or not on time by the camera person. 2) Speech transcripts do not help the segmentation process because of the significant amount of speech recognition errors in the ASR transcripts. Slides helps but depends on nature of the class and instructional styles. For instance, slides help the segmentation of lecture videos from MIS 3 because the class is more conceptual and the instructor mostly follows the slides flow in the lectures. On the other side, in MIS 2, a very hands-on programming course, the instructor spent most of his time demo programming on his laptop and explaining his codes. Most of the shots (narrative elements) in the videos from MIS 2 are demos. Thus, detecting slides change in these videos is not very helpful for segmentation although the instructor did use slides across the lecture.

Finally, we list our most important findings here.

- People use a common two step segmentation process: overall segmentation first, and refinement later.
- There are no agreements on topic structure, mostly due to granularity. People prefer linear structure to hierarchical one.
- Transitions are not necessary time points, could be fuzzy, like short time periods.
- There are no reliable segmentation features across instructors, but possible within the same instructor.
- The unprofessional or amateur nature of lecture video improves the difficulty of segmentation.
- Whether and how much external knowledge improve the segmentation depends on not only the availability of the sources, but also the nature of the class and instructional style.

With the understandings that the possible existence of common segmentation procedure and the possibility of reliable feature within one instructor, we conducted a more formal case study (discussed in the following section) to investigate the details of the segmentation procedure and potential segmentation features for one course by one instructor. We further improve the numbers of participants and videos in order to achieve more confident results.

3.2. A Case Study

The objective of our case study is to further understand the segmentation process of humans and determine the rules and heuristics that they use. More specifically, I am interested in answering the following research questions:

- What methods or processes do humans use in manual segmentation?
- What features do humans use in segmentation, and which ones do they consider to be the best?

3.2.1. Study Design

I conducted a case study with undergraduate students enrolled in an MIS course at a southwestern university. Each selected participant had earned a high mid-term score in the class. They were asked to segment lecture videos from the same class. We assume that the students' superior class performance justifies their identification as experts in the topic (besides, they had already attended the videotaped class). We also believe that their overall segmentation behavior is a good reflection of domain experts' view on segmentation.

The study consisted of two parts: a segmentation task and a questionnaire. In the segmentation task, every participant was asked to segment three lecture videos. Each lecture video was segmented by three different participants in order to check the consistency among different people. In addition, human-corrected text transcripts of the lecture video were provided to facilitate segmentation. The students were also required to

record certain statistical information such as the time they spent on the segmentation task and how many times they reviewed the video or/and transcript. After the segmentation task, they were asked to fill out a questionnaire. The questionnaire included three parts: segmentation, video/audio/transcript qualities, and general questions.

- The segmentation part included open-ended questions asking participants to identify the segmentation features they used to segment the videos. The questions were classified into five categories (video, audio, text, content and others) according to the type of input source from which the segmentation features were extracted. The "content" category was used to describe the situations when humans may only concentrate on the overall content understanding and are not aware of any specific features or cues. A list of sample segmentation features with explanations (extracted from automated segmentation research) was also provided for each category.
- The video/audio/transcript qualities part included both closed- and open-ended questions. The closed-ended questions asked participants to provide various ratings on a 7-point Likert scale. The open-ended questions asked for an explanation of their ratings of the qualities. We were interested in the impact of video, audio and text transcript qualities on segmentation because these qualities bias the results of what segmentation features are utilized by humans in the study, and thus affect our judgment on potential useful segmentation features. For instance, bad transcript quality may make it more difficult for people to extract text-based features. But it does not imply that text-based segmentation features are not good segmentation features.

- The general questions part included five open-ended questions. The participants were asked to describe their general processes in segmentation task; how they thought the video/audio/transcript qualities affected their segmentation process; and general problems and suggestions on the segmentation task.

3.2.2. Results Analysis and Findings

Thirteen participants submitted the segmentation results of eleven lecture videos and their questionnaires. Each video in the eleven videos was segmented by three out of the thirteen participants. All videos were approximately one hour (mean = 66 minutes). Average time spent on the segmentation task was 1 hour 53 minutes. After analyzing the participants' responses in the questionnaires and removing the influences of video/audio/transcript qualities, we summarized the findings as follows.

3.2.2.1. The Two-phase Process

The two-step process is further confirmed in this study. Most participants used a two-phase procedure in segmenting the videos: rough segmentation first and then refinement. Participants usually watched the video, read the associated transcript, and tried to understand the content first without marking any exact topic boundaries. Or they might simply write down some rough time stamps. Later when they reviewed the video again, they started to refine the topic boundaries and narrow down to the exact time points. Participants also claimed that they only started to notice specific indicators or cues for segmentation during refinement. For instance, some participants indicated that the

instructor will say something explicitly like “ok, let’s go to next topic…” The quantitative data also supported the idea that manual segmentation is multi-step process: most students watched the videos two or three times (with 2.66 mean review times).

3.2.2.2. Potential Segmentation Features: Combining All Input Sources

Various features from each of the sources (video, audio and text) were reported. For instance, many participants indicated that the instructor usually said some cue phrases such as “all right” or “ok” when switching to the next topic (see Figure 3.1). Shot change occurred when the instructor switched the screen from a homework review demo (topic 2) to a PowerPoint slide (topic 3). The most commonly used features reported by participants are listed in Table 3.2.

Topic 1: Introduction to Lecture Content
“*All right*, today we get to do the funner stuff. This is where we get into the whole purpose behind java, so today is gonna be good. Object oriented programming, that is what we are all here for …”

Topic 2: Homework review-two-dimensional arrays
“*Ok* but first of course we need to go through the homework assignment that I assigned for this time, which was to write an assignment called print2darray that accepts a two dimensional string array as a parameter …”

Topic 3: Object Oriented Programming Lecture
“*All right*, so that was that. Alright, object oriented programming. We are finally in the real meat. This is the real whole purpose behind java the fact even exist followed up the object oriented programming …”

Figure 3.1. Part of the transcript for a lecture video about Java Programming.
NOTE: The “Topic” headers indicate the boundaries identified by study participants. Certain words or phrases are bolded to show the segmentation features used.

However, we found that no single source or universal feature was used by all of the participants across the segmentation process. Instead, participants made use of features from all three sources: video, audio and text. For instance, in Figure 3.1 while judging the topic transition from topic 2 to topic 3, people made use of features from all three sources: the shot change from the video, the cue phrase "all right" from the text transcript and the high pitch from the audio when speaking the "all right." This implies that humans use complementary features from all three sources in order to make a decision on a topic boundary. Furthermore, several students also reported that they found the lecture slides to be very helpful in the segmentation process.

Table 3.2. Potential segmentation features identified in manual segmentation

Category	Potential Segmentation Features
Video	*Scene/shot changes:* "When the instructor is explaining about java theory, the camera also moves to the white board - when the instructor is doing the homework problem, the camera will be on the slides" (problem: course and instructor style specific) *Gesture/posture:* "Turn around to whiteboard or erase the whiteboard"
Audio	*Long pause or silence:* "Instructor usually pauses before changing topic" *High pitch:* "Alright!", "Yes!" with higher pitch
Text	*Cue phrases:* "Ok so let's go on to…", "Now we are going to talk about….", "Instructor usually said explicitly that he wanted to move to next topic" *Introduction of new vocabulary:* "the instructor stated the topic prior to talking about the new topic"; new topic words were discussed (e.g. "composition", "inheritance", etc) before going to that topic.
Content	*Overall content changes:* Comments such as "where the instructor talks about another topic"; "When he started to talk about something other than what he was talking about earlier" are very common
Others	*Questions:* Topics were changed by students' questions when the instructors asked students if they had further questions on the current topic.

3.3. Conclusion

We conducted two studies on manual segmentation to acquire more understanding of the topic structure of a lecture video, and collected features and rules to facilitate the design of automated method. The major findings of two studies and their implications on developing the automated segmentation method are summarized as follows.

- *Content Structure*: The first finding of content structure is that a linear structure is generally more preferred than a hierarchical one. Thus, our automated segmentation research concentrates on methods that segment a lecture video into a linear list of topics. The second finding of "topic transitions could be short time periods as well as exact time points" indicates that a topic boundary could be fuzzy and certain degree of boundary relaxation should be allowed during evaluation.
- *Segmentation Features*: Because there are no reliable segmentation features across courses/instructors (but possible for one course/instructor) except the overall content changes, we focus on developing methods which can capture the overall content changes, and general methodology that are adaptive or can be easily customized to various courses/instructors. Our first text-based segmentation algorithm is such a method that can capture the overall content change across topics because speech text carries most of information a lecture video. On the other hand, the second multi-modal we proposed is more a methodology rather than a specific method. It includes the general framework and principles to integrate features from multiple modalities instead of certain specific features. A set of potential segmentation features and their

extraction methods are summarized in Table 3.3 for scholars who are interested in developing an automated segmentation approach for lecture videos.

- *Segmentation Process*: One major finding in the two studies is the common two phase segmentation process: rough segmentation and segmentation refinement. Content based features such as color histogram and term frequency are usually good for identifying rough boundaries at the first step; while discourse based features such as pause length and cue phrases are good for refining (or confirming/rejecting) the rough boundaries. The framework in the proposed multimodal approach in Chapter 4 incorporates the idea of two phase process.

Extra Knowledge: extra knowledge sources such as slides, if available, are no doubt an aid on segmentation. However, whether and how much extra knowledge can improve the segmentation depends on not only the video types (e.g. some instructor may do not use slides in classes), but also on the nature of the class and the teacher's instructional style. For instance, in one lecture/class the instructor follows the flow of the slides strictly. While in another lecture or class, s/he may rely on blackboard and use the slides occasionally for further illustration purpose. This further confirms that a method fully relies on slide matching is not enough for our topic segmentation purpose (Section 2.4, Chapter 2). Other types of information from speech text or audio (e.g. prosodic cues) should be used with them to complement each to achieve the best performance. The method we used in the second preliminary results of our multimodal approach is a good example (Section 4.2, Chapter 4).

Table 3.3 Summary of segmentation features and extraction methods

Input Source	Phase	Segmentation Features		Extraction Method	Segmentation Level
Image/Video	Content-based	Scene/shot changes (color histogram, edge, motion, statistical analysis) Number & size of faces		• Color histogram distance computation between successive images (Wactlar 2000) • Progressive transition detection by combining both motion and statistical analysis (Zhang and Smoliar 1994) • Face detection (Viola and Jones 2001) (Chaisorn et al 2003)	Narrative segmentation
		Color histogram, slide content area changes, key points, video text recognition, topic event detection		• Detect local color histogram differences (Li and Dong 2006) • Extract key points from frames and slide images (Fan et al 2006) • Topic video event detection (Syeda-Mahmood and Srinivasan 2000)	Slide Matching
	Discourse-based	Hand gestures, Head Pose, Posture shift		• Skin color detection and HMM for gesture recognition (Wang et al 2004 & 2005) (Martin and Durand 2000)	Topic Segmentation
Audio	Discourse-based	Phone and rhyme duration		• Speech recognizer output (Shriberg et al 2000)	
		Pause duration		• Speech recognizer output (Shriberg et al 2000)	
		Pitch		• Pitch tracker (F0 features output) (ESPS 1993) (Shriberg et al 2000)	
		Speaker turns or gender changes		• Speech recognizer output (Shriberg et al 2000)	
Text	Content-based	Introduction of new vocabulary (e.g. words or noun phrases)		• Vocabulary Management Profiles (Youmans 1991) • Natural language processing (tokenization, sentence splitting, POS tagging, word stemming, noun phrase extraction, and co-occurrence analysis) on sentences transcribed from the speech in video. • (Lin et al., 2005).	
		Overall content changes	Term (e.g. word, noun phrase) frequency		
			Lexical chain	• Term (e.g. noun phrases, synonyms, name entities, co-references) repetition (Kan et al., 1998)	
			Graphic representation	• Dotplotting (Reynar 1994) (Choi 2000)	
	Discourse-based	Pronouns		• Mapping to a list of pronouns (Lin et al 2005)	
		Cue phrases		• Mapping to a list of common cue phrases (Lin et al 2005) • Machine learning (sparse training data)	
		Questions		• Sentence structure analysis to identify questions	

CHAPTER 4

AUTOMATED STATIC SEGMENTATION METHODS

As research challenge remains on recognizing and extracting the topic structures of lecture videos, this chapter presents two segmentation algorithms which segment a lecture video into topics. While the manual segmentation study shows that a linear structure rather than hierarchical one is preferred by human, both algorithms segment a lecture video into a linear list of topics. The first algorithm is a text based approach that uses natural language processing techniques such as noun phrases extraction and lexical knowledge sources such as WordNet. Multiple linguistic based segmentation features (e.g. such as noun phrases and cue phrases) are extracted from the speech transcript accompanying with the lecture videos. The second algorithm is a multi-modal method designed based on the results and findings from the manual segmentation studies (Chapter 3). The algorithm combines segmentation features from three information sources of video (speech text transcript, audio and video) and makes use of various knowledge sources such as world knowledge and domain knowledge. It also simulates the two phase process commonly used in manual segmentation: initial segmentation and segmentation refinement. The experiment results indicate that noun phrase are salient features and both methods are promising in topic segmentation of lecture videos.

The chapter is organized as follows. Section 4.1 introduces the text based segmentation and its evaluation. Section 4.2 presents the multi-modal approach and the preliminary results. Section 4.3 concludes the chapter by summarizing the contributions and pointing out the future directions.

4.1. A Text Based Segmentation Approach

Since audio, especially the speech in the audio conveys most of the information in a lecture video, my first algorithm concentrate on segmenting the speech text transcript extracted from the audio track. Speech text is extracted automatically using speech recognition software and corrected by human. The segmentation results on the speech text can be mapped back to video using the time codes associated with the speech text. The segmentation task is, therefore, to automatically segment the text transcripts into topically cohesive blocks by detecting topic boundaries.

However, as indicated in Chapter 2, unlike the segmentation methods that focus on written text, segmentation of transcribed spoken text is more challenging because spoken text lacks typographic cues such as headers, paragraphs, punctuation, or capitalized letters. Moreover, compared to written text and news stories, the topic boundaries within lecture transcripts tend to be more subtle and fuzzy because of the unprofessional and spontaneous speech and the large variety of instructional methods. More resolving power is required for segmenting lecture transcripts. With the advancement of computational linguistics and natural language processing (NLP) research, NLP techniques such as Part-

of-Speech tagging and noun phrase extraction are becoming more mature and available for real life usage. They are potentially useful for gaining more resolving power and improving segmentation accuracy because they provide a deeper structure and better understanding of the language and content in the transcript. We propose a linguistics-based approach which utilizes all kinds of linguistics-based features and NLP techniques to facilitate the automated segmentation. Part-of-speech tagging was used to distinguish between different word types (noun, verb, adjective). Noun-phrase extraction could help the segmentation because noun phrases carry more content information than single words (Katz 1996). Lexical knowledge such as WordNet was used because different words such as synonyms may be used to express the same concept in a text. The basic idea behind the proposed approach is that different linguistic units and features (e.g. different word types, larger units such as noun phrases, discourse markers or cue phrases) carry different portions of content and structure information and therefore should be assigned different weights. More specifically, I am interested in two research questions as follows.

(1) As the names of concepts and theories that appear frequently in lectures are usually noun phrases, are noun phrases more salient segmentation features and could they be used to improve segmentation performance?

(2) Intuitively, linguistic features modeling different characteristics of text (e.g. content-based vs. discourse based) should complement each other, then can the combination of multiple linguistic segmentation features lead to gains in resolving power and thus improve segmentation performance?

4.1.1 The Approach

To answer the above research questions, I propose implementing an algorithm called PowerSeg. The algorithm combines multiple linguistic segmentation features which include content-based features such as noun phrases and verbs, discourse based features such as pronouns and cue phrases. It also incorporates lexical knowledge from WordNet to improve accuracy. The algorithm utilizes an idea similar to the sliding window methods in TextTiling (Hearst, 1994). We move a sliding window (e.g. *W*1, *W*2) of certain size (e.g. 6 sentences) across the transcript by certain interval (e.g. 2 sentences) (Figure 4.1). We then compare the similarities between two neighboring windows of text. For instance, we compute the similarity between windows *W*1 and *W*2 (e.g. sentence numbers 14-19 vs. 20-25); then we move *W*1 and *W*2 by an interval of 2 sentences, and calculate the similarity between *W*1(2) and *W*2(2). I repeated this process until the sliding windows reach the end of the transcript. The places where similarities have a large variation are identified as potential topic boundaries. The basic idea here is that I view the task of topic-based segmentation as the detection of shift from one topic to the next. In other words, the task is to detect where the use of one set of terms ends and another set begins (Halliday & Hasan, 1976). Then the remaining questions are how we calculate the similarities between two windows, how we represent the topic information of each sliding window, and finally how we identify the largest variations of similarities. We will answer all these questions in the detailed algorithm description. Basically the algorithm has three major steps: (1) Preprocessing. (2) Features extraction. (3) Finding boundaries.

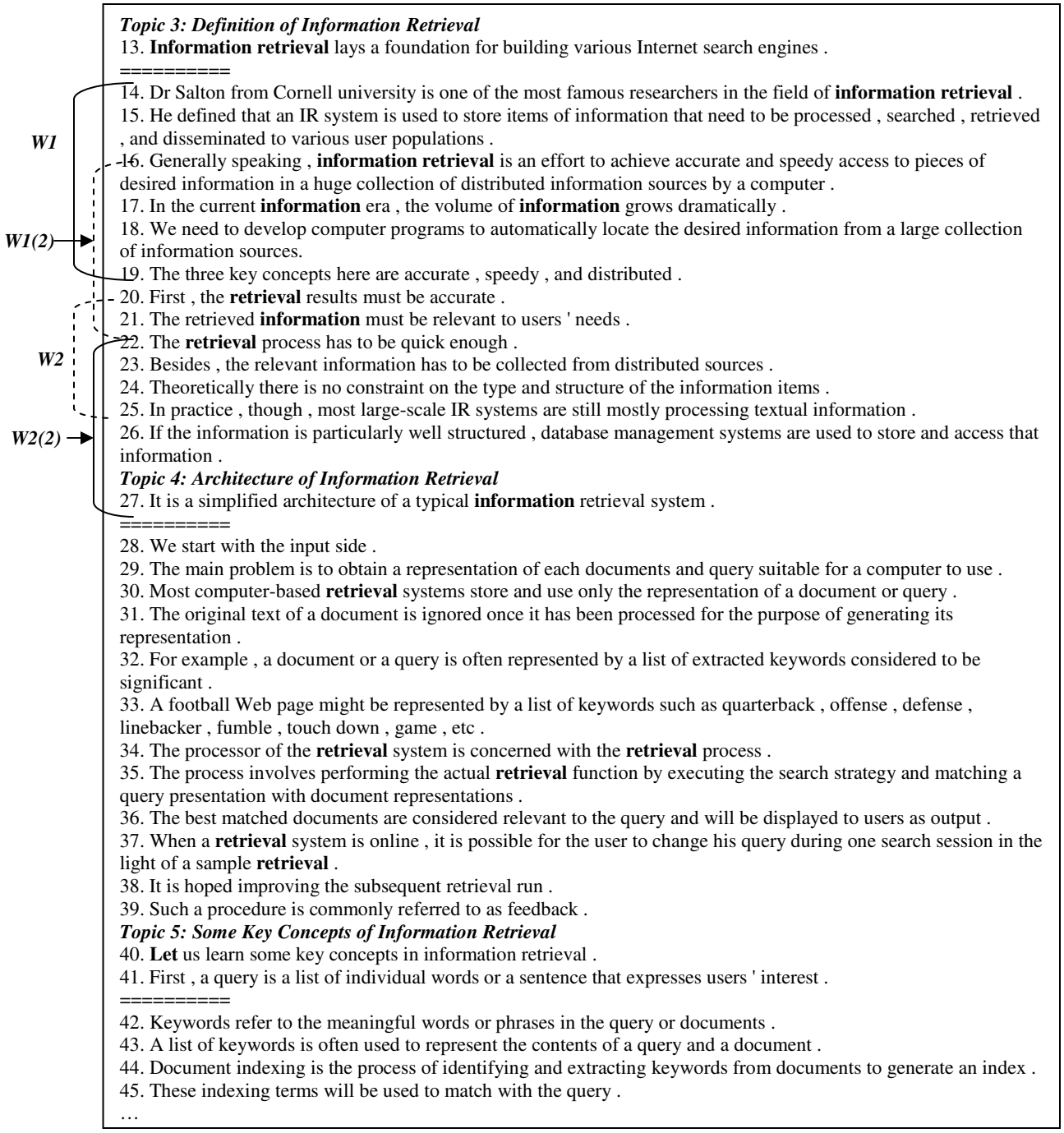


Figure 4.1. Part of the transcript for a lecture video about Information Retrieval. Each line starts with the sentence number. Lines with "==========" are boundaries identified by automated algorithm. Lines start with "Topic:" are actual boundaries identified by human experts.

The preprocessing step performs preparation work for next steps, which is literally standardized. The algorithm takes the transcript text as input, and uses GATE (Cunningham 2000) to handle tokenization, sentence splitting, and part-of-speech (POS) tagging. GATE is a widely used human language processing system developed at the University of Sheffield. GATE splits the text into simple tokens such as numbers, punctuation and words, segment the text into sentences, and the part-of-speech tag was produced as an annotation on each word or symbol e.g. NN for nouns and VB for verbs). Further, Porter's stemmer (Porter 1980) is used for suffix stripping (e.g. "lays" becomes "lay"). Punctuations and uninformative words are removed using a stopword list. Based on the results from preprocessing, different features, such as noun phrases, are extracted to represent each sliding window and used for similarities comparison.

4.1.1.1. Feature Extraction

Seven feature vectors are extracted including noun phrases (NP), verb classes (VC), word stems (WS), topic words (TNP), combined features (NV), pronouns (PN), and cue phrases (CP). The first five features (NP, VC, WS, TNP and NV) are content-based features, which carry lexical or syntactic meanings of the body of content. The last two features (PN and CP) are discourse-based features, which describe more about the properties of the small text body surrounding the topic boundaries.

We use noun phrases instead of “bag of words” (single words) because noun phrases are usually more salient features and exhibit fewer sense ambiguities. Furthermore, most names of concepts and theories in a lecture are noun phrases. For instance, in the transcript of a lecture video about Web search engines (see Figure 4.1), topic 3, “Definition of Information Retrieval” and topic 4, “Architecture of Information Retrieval” share a lot of words such as “information” and “retrieval” (in bold face in Figure 4.1). It will be hard for algorithms using single-word features such as word repetition to distinguish between these two topics. However, it will be much easier to separate these two topics if we use noun phrases. For instance, “information retrieval” occurs several times in topic 3, but not in topic 3. We use the Arizona Noun Phraser (Tolle and Chen 2000) to extract the noun phrases from transcript.

Besides noun phrases, verbs also carry a lot of content information. Semantic verb classification has been used to characterize document type (Klavans and Kan 1998) because verbs typically embody an event’s profile. Our intuition is that verb classification also represents topic information. After removing support verbs (e.g. is, have, get, go, etc., which do not carry a lot of content information), we use WordNet to build the links between verbs to provide a verb-based semantic profile for each text window during the segmentation process. WordNet is a lexical knowledge resource in which words are organized into synonym sets (Miller et al 1990). These synonym sets, or synsets, are connected by semantic relationships such as hypernymy or antonymy. We use the synonymy and hypernymy relationship within two levels in WordNet. We only accept

hypernymy relationships within two levels because of the flat nature of verb hierarchy in WordNet (Klavans and Kan 1998). More specifically, when two verbs between two text windows are compared, they will be considered as having the same meaning (or in the same verb class) if they are synonyms or hypernyms within two levels. Except nouns and verbs, other content words such as adjectives and adverbs will be simply used in their stem forms (word stems, WS).

Other than those simple features (nouns, verbs and word stems), we also have two complex features. The first one is topic terms, or more exactly, topic noun phrases. Topic terms are defined as those terms with co-occurrence larger than one (Katz, 1996). Topic terms usually hold more content information (such as "information retrieval" in Figure 4.1), which means they should carry more weight in our algorithm. The other complex feature is a combined feature of nouns and verbs. We extract the main noun and verb in each sentence according to the POS tags, with the expectation of capturing the complex relationship information of subject plus behavior.

Different from the above five content-based features, the two discourse-based features focus on the small size text body surrounding the pseudo-boundaries proposed by the algorithm based on the five content-based features. We use a size of five words in our algorithm. In other words, we check the five words before and after the pseudo-boundaries. If we find any pronoun (from a pronoun list) within the five-word window, we decrease the probability score of this pseudo-boundary as a true boundary. The reason

is that pronouns usually substitute for nouns or noun phrases that appear within the same topic. Any occurrence of cue phrases (from a cue phrase list) will increase the probability of pseudo-boundary as a true boundary because cue phrases usually indicate the change of discourse structure (e.g. cue phrase "Let" at the beginning of topic 5, Figure 4.1.). We use the general cue phrases list (Table 4.1) and the pronoun list (Table 4.2) from Reynar (1998).

Table 4.1. Cue phrases

actually	further	otherwise
also	furthermore	Right
although	generally	Say
and	however	Second
basically	indeed	See
because	like	Similarly
but	look	Since
essentially	next	So
except	no	Then
finally	now	Therefore
first	ok	Well
firstly	or	Yes

Table 4.2. Pronouns

she
her
hers
herself
he
him
his
himself
they
their
them
theirs
themselves

4.1.1.2. Similarity Measure

The similarity between two neighboring text windows ($w1$ and $w2$) is calculated according to cosine measure in vector space model (Salton et al 1989). Given two neighboring text windows, their similarity score is the sum of normalized inner product of seven feature vectors weights. The basic idea is that neighboring text windows with more overlapping of features (e.g. noun phrases, verbs) will have higher similarity.

$$Similarity(w_1, w_2) = \sum_j \frac{\sum_i f_{j,i,w1} f_{j,i,w2}}{\sqrt{\sum_i f^2{}_{j,i,w1} \sum_i f^2{}_{j,i,w2}}} S_j \quad (4.1)$$

j represents the different features (1 to 7 here), and *i* ranges over all the specific feature weight values (e.g. noun phrases) in the text window. *fj,i,w1* is the *i*-th feature weight value of *j*-th type feature vector in text window *w*1. We calculate *fj,i,w1* based on term frequency (TF). *j* is the feature type and *i* is the specific word or noun phrase in the feature vector. *Sj* is the significant value of some specific feature type. The best way to calculate *Sj* is to use language model or word model and utilize large corpus. For instance, Reynar (23) uses G-model and Wall Street Journal to calculate *Sj* (called word frequency in Reynar (1998)). However, without large training corpus of lecture videos available, the significant values *Sj* are estimated based on human heuristics and hand tuning. We assume that significances of the five features are in the following order: *S* (TNP) > *S* (NV) > *S* (NP) > *S* (VC) > *S* (WS).

4.1.1.3. Finding the Boundaries

After the similarity between two neighboring windows for each interval is calculated, a similarity graph for all the intervals is drawn (see Figure 4.2). Intervals are certain number locations in the text transcript (e.g. 13, 15, 17 …), similar to the concept of "gap" in TextTiling (Hearst, 1994). The X-axis indicates intervals and Y-axis indicates similarity between neighboring windows at each corresponding interval (e.g. the interval at sentence number 17). The intervals with largest depth values (deepest valleys) are identified as possible topic boundaries. The depth value is based on the distances from

the valley to the peaks to both sides of the valley, which reveals how strongly the features of topics (e.g. frequency of noun phrases occurring) change on both sides. For instance, the depth value of the interval at sentence number 27 is equal to (y3 - y2) + (y1 - y2) (see Figure 4.2). To decide how many boundaries the algorithm will assign, we use a cutoff function ($m - sd$). m is the mean of all depth values and sd is the standard deviation. In other words, we draw boundaries only if the depth values are larger than ($m - sd$).

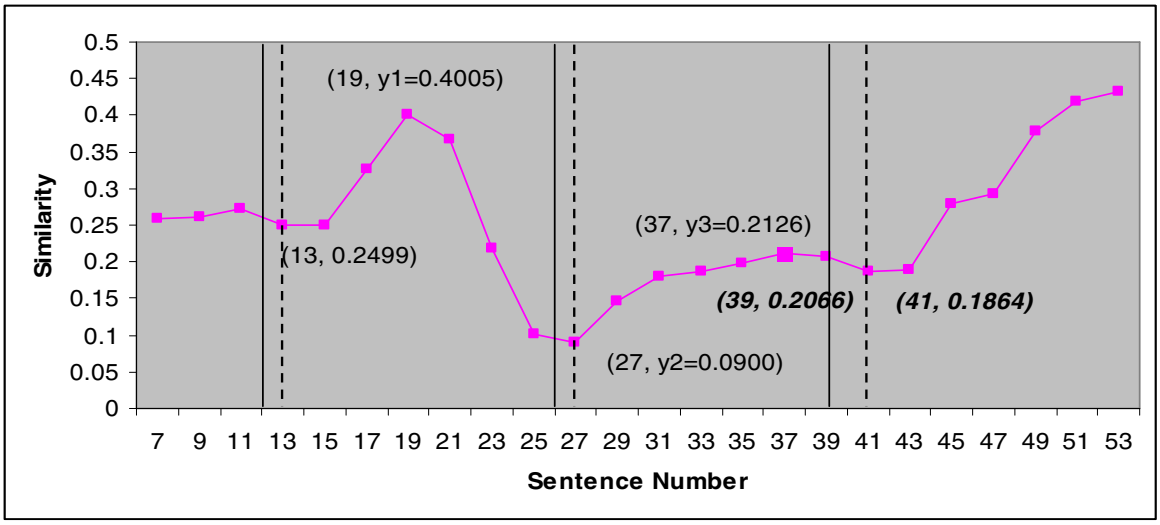


Figure 4.2. Example of a similarity graph. Dashed vertical lines indicate the boundaries proposed by automated method (e.g. PowerSeg here). Solid vertical lines indicate the actual boundaries.

4.1.2. Evaluation

To test our research questions that noun phrases are salient features and that the combination of features improve accuracy, we evaluated our algorithm with a subset of features. We chose five features (NP, TNP, WS, CP, PN) to conduct a preliminary experiment. Those five features include salient features such as noun phrases (NP) and a

combination of both content- and discourse-based features: NP, TNP (topic noun phrases) and WS (word stems) for content-based features and CP (cue phrases) and PN (pronoun) for discourse-based features. The performance of PowerSeg was compared to that of a baseline method and TextTiling (Hearst 1994), one of the best text segmentation algorithms. For TextTiling we used a Java implementation from Choi (2000). We also developed a simple version of the baseline segmentation algorithm. The baseline algorithm randomly chose points (e.g. certain sentence numbers) to be topic boundaries. We hypothesized that:

H1: *The PowerSeg algorithm with NP alone achieves a higher performance than TextTiling and Baseline.*

H2: *The PowerSeg algorithm with NP+CP+PN achieves a higher performance than PowerSeg with NP alone or WS alone.*

4.1.2.1 Data Set and Performance Metrics

Since there was no available annotated corpus for lectures videos, we used the lecture videos in our e-Learning system called Learning By Asking (LBA) (Zhang 2002) as pilot data for evaluation. Because the task of transcribing lecture videos is very time consuming, we choose a small data set of three videos for our preliminary experiment. All three videos are chosen randomly and transcribed by human experts. The three videos are selected from two different courses and instructors. One video was from a lecture about the Internet and Web search engines, and the other two were from a database course. Three transcripts corresponding to the videos were used for the evaluation

purpose. The average length of the videos is around 28 minutes and the average number of words in the transcripts is 1,859. We assumed that the segmentation results from the experts are perfect (100% accuracy). The performance measures of PowerSeg, TextTiling, and Baseline were calculated by comparing their output results to the results from the experts.

Selecting an appropriate performance measure for our purpose is difficult. The metric suggested by Beeferman et al (1997) is well accepted and has been adopted by TDT. It measures the probability that two sentences drawn at random from a corpus are correctly classified as to whether they belong to the same story. However, this metric cannot fulfill our purpose because it requires some knowledge of the whole collection and it is not clear how to combine the scores from probabilistic metrics when segmenting collections of texts in different files (Reynar 1998). Instead, we chose precision, recall and F-measure as our metrics. Precision and recall were chosen because they are well accepted and frequently used in information retrieval and text segmentation literature (Hearst, 1994; Reynar, 1998). F-measure was chosen to overcome the tuning effects of precision and recall. Precision, recall and F-measure were defined as follows:

$$\mathrm{P(recision)} = \frac{\mathrm{No\_of\_Matched\_Boundaries}}{\mathrm{No\_\ of\_Hypothesized\_Boundaries}} \qquad \mathrm{R(ecall)} = \frac{\mathrm{No\_of\_Matched\_Boundaries}}{\mathrm{No\_\ of\_Actual\_Boundaries}}$$

$$\mathrm{F\text{-}Measure} = \frac{2PR}{P+R}$$

No_of_Matched_Boundaries is the number of correctly matched boundaries when comparing to actual boundaries identified by experts. No_of_Hypothesized_Boundaries

is the number of boundaries proposed by the algorithm (e.g. PowerSeg). Besides exact match, we also used the concept of fuzzy boundary which means that hypothesized boundaries that are a few sentences (usually one) away from the actual boundaries are also considered as correct. We used fuzzy boundary because for lengthy lecture videos, one sentence away from the actual boundary is acceptable for most applications. It is only a very short time period when we map the transcript back to the video. For instance, the average time span of one sentence in our data set is only 12 seconds.

4.1.2.2. Experiments and Results

We ran the three algorithms (Baseline, TextTiling and PowerSeg) using the three transcripts and calculated the mean performance. The performance measures (precision, recall and F-Measure) were calculated under two conditions: exact match and fuzzy boundary. Under fuzzy boundary condition, hypothesized boundary that is one sentence away from the actual boundary is acceptable.

4.1.2.2.1. Hypothesis Testing

First, in order to test whether noun phrases are salient features (H1), we ran the PowerSeg algorithm with the NP feature only, TextTiling and baseline using our dataset. We found that even with NP only, PowerSeg improved the performance (F-Measure) by more than 10% compared to both Baseline and TextTiling under “fuzzy boundary” condition (Table 4.3). Under the “exact match” condition, the PowerSeg only performed 5% better than Baseline, although it was 15% better than TextTiling. Surprisingly, the TextTiling

algorithm performed even worse than Baseline. It showed that the "bag of words" algorithms were not good at identifying exact topic boundaries especially when the transcript is about sub-topics with a lot of words shared (e.g. topics 13 and 14 in Figure 4.2). On the other hand, all algorithms performed better under the "fuzzy boundary" condition as expected.

Table 4.3. Comparison of algorithm

Algorithms	Exact Match			Fuzzy (1)		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Baseline	0.32	0.32	0.32	0.56	0.56	0.56
TextTiling	0.30	0.18	0.22	0.75	0.46	0.56
PowerSeg(NP)	0.41	0.35	0.37	0.77	0.67	0.70

In order to evaluate the effectiveness of features combination (H2), we ran four different versions of PowerSeg which used 4 types of feature subsets: WS (word stem only), NP (noun phrase only), NP+TNP (noun phrase plus topic noun phrases) and NP+CP+PN (noun phrases, cue phrases, and pronouns) (Table 4.4). We found that the combination of noun phrases, cue phrases, and pronouns had a better performance than using noun phrases only (NP). This showed that the combination of multiple features, especially the combination of content-based features and discourse-based features, improved segmentation performance (F-Measure).

Table 4.4. Comparison of PowerSeg with different feature subsets

Features	Exact Match			Fuzzy (1)		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Baseline	0.32	0.32	0.32	0.56	0.56	0.56
WS	0.30	0.18	0.22	0.75	0.46	0.56
NP	0.41	0.35	0.37	0.77	0.67	0.70
NP+TNP	0.39	0.32	0.34	0.73	0.60	0.65
NP+CP+PN	0.42	0.37	0.39	0.77	0.68	0.72

However, the improvement was very small, only around 2%. The possible reason was that the cue phrase list and pronoun list we used are too general given our small data set. Those words may happen rarely in the small dataset. To our surprise, the NP+TNP combination performed slightly worse than using NP only. One possible reason is that although we defined topic noun phrases as those noun phrases with frequency larger than one, our feature weighting method and calculation of similarity were still based on term frequency. When we calculated the similarity between two text windows, TNPs already occupied a large percentage of weight. From another perspective, it also showed that complementary features such as content-based features and discourse-based features would improve performance, but not those with similar characteristics such as noun phrases and topic noun phrases.

In summary, H1 has been supported but H2 has not. In other words, the PowerSeg algorithm using noun phrases alone performed better than the Baseline and the TextTiling methods. Referring to our first research question, we have shown that noun phrases are

salient linguistic features that can greatly improve the performance of video segmentation systems. We suggest that noun phrasing can be useful for video indexing and other applications. However, concerning the second research question, we found that the combination of different linguistic features did not further improve the performance of the algorithm. Beside small test dataset and general cue phrase and pronoun list, another possible reason is that noun phrases are very important and already represent most of the topic and content information in the text. Therefore, the addition of other features does not provide more useful information to the algorithm.

4.1.2.3. Discussion

Overall the experiment results are promising. Our proposed PowerSeg algorithm achieved 0.70 in F-measure when noun phrases were used as the only feature and the fuzzy boundary was applied. As it has been shown that human agreement on video segmentation is often only around 0.76 (Precision: 0.81; Recall: 0.71) (Hearst 1994), our algorithm has performed similarly by agreeing well with the segmentation generated by our human experts.

Because of the distinct characteristics of datasets and different performance measures (as described in our literature review and in evaluation sections), it is hard to compare the segmentation results with those achieved in other domains such as broadcast news segmentation. However, the segmentation of lecture videos is expected to be more difficult because of the lack of large training datasets and the large variety of instructional

methods. For instance, the formal presentation format and cue phrases that the methods in the news domain heavily rely on are not available for lecture videos. As previous research shows that the segmentation performance of the HUB-4 news broadcast data, measured by precision and recall, is only around 0.6 (Reynar 1998), our algorithm achieves a promising performance. For further comparison, future research needs to be conducted to evaluate the performance of our algorithm using broadcast news data.

4.2. A Multimodal Segmentation Approach

The second automated video segmentation algorithm we developed is a multi-modal method which combines features from multiple modalities. Furthermore, it is based on the findings from our manual segmentation study and related research of automated segmentation. The proposed approach has several novelties: it 1) simulates the two-phase manual segmentation process with initial segmentation and segmentation refinement, 2) combines segmentation features from all three input sources, and 3) utilizes various knowledge sources including world knowledge, domain knowledge and extra knowledge. As shown in Figure 4.3, the proposed automated segmentation method has three steps: preprocessing, feature extraction, and features fusion and segmentation. Details are discussed as follow.

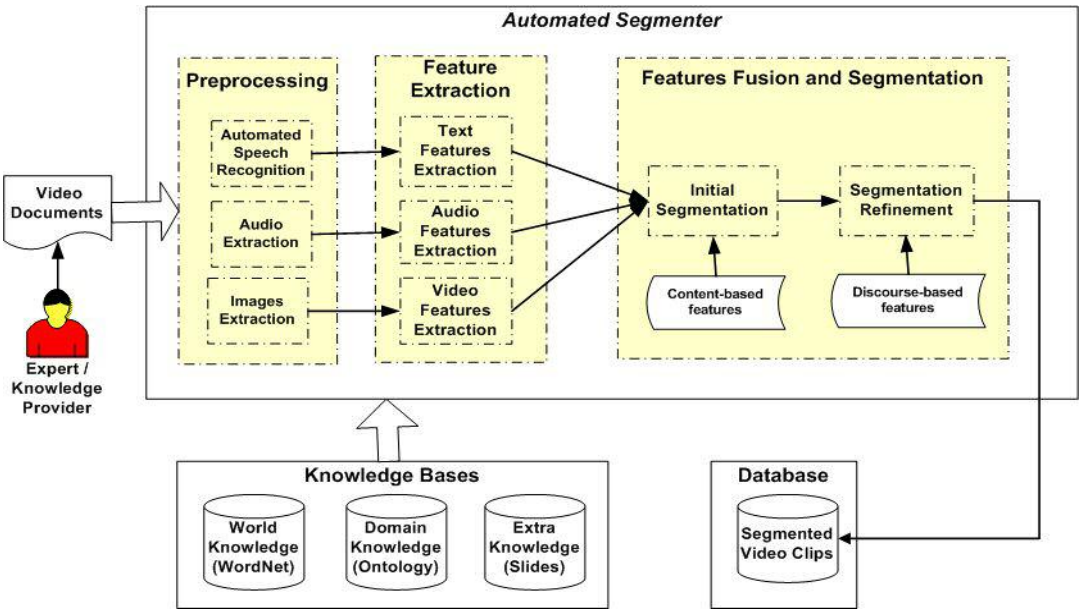


Figure 4.3. A Multimodal segmentation method

4.2.1. The Detailed Algorithm

The preprocessing step processes raw video and prepares for feature extraction. Text transcripts are extracted using automated speech recognition (ASR) software. Image sequences and audio are also extracted. In feature extraction, various features are extracted, including text-based features (such as noun phrases and cue phrases), and low level video and audio features such as color histogram. All features we identified in the manual segmentation study could be potential candidate features (Chapter 3, Table 3.2). Finally all features are fused and used for segmentation in the features fusion and segmentation step. Knowledge sources are used across the segmentation process.

In the features fusion and segmentation step, all features are passed to a two-phase segmentation process: initial segmentation and segmentation refinement. Initial

segmentation performs a rough segmentation using content-based features. Because the speech in a lecture video contains the majority of the content information, most of the content-based features are extracted from speech-transcribed text. These features are usually linguistic features such as noun phrases, verb classes and word stems, which have already been identified as salient features in our previous study (Lin et al 2005). Another reason we focus on the text transcript is the low computational cost of text processing compared with video/image and audio processing. After identifying potential topic boundaries in initial segmentation, we refine the rough boundaries using more computationally expensive features extracted from video and audio in segmentation refinement. Differing from content-based features used in initial segmentation, most features used in segmentation refinement are discourse-based features. They describe the characteristics of the small body of content immediately adjacent to the potential boundaries proposed in initial segmentation. For instance, most of the segmentation features we identified in manual segmentation belong to discourse-based features including video features (e.g., shot changes, gestures and posture shifts), audio features (e.g., pause length and pitch) and text features (e.g., cue phrase and introduction of new words). Furthermore, the computational cost can be significantly decreased when those computationally expensive features are only calculated in the small windows around the potential boundaries.

Although the combination of segmentation features from the three sources is beneficial, the fusion of these features is not an easy task. One possibility is to adapt the sliding

window method used in previous research (Lin et al 2005). After identifying potential topic boundaries (by finding the points with the most dissimilar neighboring windows) in initial segmentation, features applied in refinement are only used to increase or decrease the probability of the potential boundaries as actual boundaries. However, if a training corpus was available, machine learning methods will be more reliable and achieve better performance. One method is to use a decision tree or maximum entropy model (Reynar 1998). The decision of whether or not the initial segmentation suggests a topic boundary (indicated by a 1 or 0) is one feature that is used (along with all other features) in the refinement. Since segmentation features from different sources have a variety of characteristics, a combination of a decision tree and a Hidden Markov Model (HMM) may be a better strategy. The posterior probability from HMM using text features in initial segmentation will be used as a feature in the decision tree with all video and audio features. We can also use HMM as a top-level model (Shriberg et al 2000).

4.2.2. Knowledge Bases

During the entire segmentation process, various knowledge bases are utilized to assist in the segmentation. World knowledge includes knowledge sources about the general world including sources such as WordNet. WordNet is a lexical knowledge base in which words are organized into synonym sets (Miller et al 1990). These synonym sets, or synsets, are connected by semantic relationships such as hypernymy or antonymy. WordNet has been used to extract the “verb class” feature in previous research (Lin et al 2005). Domain knowledge includes ontology extracted from sources such as the Internet, electronic

textbooks or professional dictionaries. Because many concepts in a lecture are professional terms in a specific domain, they do not exist in a general lexical knowledge base such as WordNet. However, we can extract relationships from, for instance, a professional dictionary (e.g., Webopedia, a dictionary for computer and Internet technology definition) and build a WordNet-like ontology. In many situations, extra knowledge such as an instructor-created lecture outline and slides are also available for lecture videos. They provide valuable information for the segmentation. For instance, slides have been used to correct the word errors in the ASR transcripts (Cao 2004).

4.2.3. Preliminary Results

We conducted a preliminary evaluation using the same method from previous research (Lin et al., 2005). Previous method makes use of content-based text features in the first stage and discourse-based text features in the second stage, which is a good match to the proposed two-phase segmentation process although only text-based features were used. We retested the algorithm using a different dataset by randomly selecting three lecture videos from two different MIS courses and ensuring they were different videos than previously used (in Lin et al 2005). We focused on evaluating the effectiveness of the two phase segmentation: initial segmentation and segmentation refinement. We achieved similar results to what was previously found (in Lin et al 2005) (Table 4.5). We found a small improvement between the algorithm implementing the initial segmentation step only (“Initial”) and the algorithm implementing both steps (“Initial + Refinement”) (1.6% in F-Measure and 4.0% in Precision). However, the improvement is not significant. One

possible reason could be that the cue phrases and pronouns we used (Lin et al 2005) are too general and may happen rarely in our small test dataset. Another reason could be that all features used in (Lin et al 2005) come from the text source only. It is expected that the incorporation of video and audio features as proposed in this paper will complement the text features and achieve better performance levels as a result.

Table 4.5. Evaluation of the effectiveness of automated segmentation method

Version	Exact Match			Fuzzy (1)		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Initial	0.35	0.25	0.29	0.75	0.56	0.64
Initial + Refinement	0.37	0.25	0.30	0.78	0.56	0.65

4.3. Conclusion and Future Directions

Two automated segmentation algorithms were developed in order to extract the topic structure of lecture videos. The first topic segmentation algorithm makes use of multiple linguistic features extracted from the speech transcripts. It utilizes salient linguistic segmentation features such as noun phrases, and combined content-based and discourse-based features to gain more resolving power. Experiment results demonstrated that the effectiveness of noun phrases as salient features and the methodology of combining multiple linguistic based text features to complement each other are promising. The algorithm achieves an accuracy of 0.70 in F-Measure. The second algorithm is a multimodal approach which incorporates features from video, audio and speech text, adapts a two-phase segmentation process (initial segmentation plus refinement) from

manual segmentation studies, and utilizes various knowledge bases. The preliminary demonstrate the methodology is promising.

There are several limitations of our algorithms. First, although the evaluation results of both algorithms are encouraging, one should note that relatively small datasets were used in our experiment. Caution needs to be taken when interpreting our findings. More evaluations on larger sets of data will be needed to increase the reliability and validity of our results. Therefore, one of future direction is to test our algorithms on large data sets from different instructors and courses. Second, the "transcript problem" needs to be addressed. The performance of the video segmentation algorithm depends greatly on the correctness of the transcripts. Currently, when transcripts of the videos are not available, they have to be created using speech recognition software, which often does not achieve high accuracy. Such transcripts have to be corrected manually in order to ensure better performance in video segmentation. Our dynamic segmentation approach (Chapter 5) addresses this problem partially by correcting the speech transcripts automatically using external knowledge sources such as electronic slides. Third, the method is currently designed and tested for English lectures only. Noun phrases, while salient in English, may not be as useful in other languages. Some components in our system are also language-specific (e.g., the speech recognition software). Customization of the system, tuning of the parameters, and further testing will be necessary if the algorithm is applied to videos in a language other than English.

CHAPTER 5

A QUERY SPECIFIC SEGMENTATION APPROACH

This chapter presents a novel approach to segment lecture videos into relevant segments and return them to users based on specific questions asked by users. In the proposed approach, shallow parsing such as part of-speech and noun phrase chunking were used to parse both questions and Automated Speech Recognition (ASR) transcripts. A sliding window approach was used to identify the start and ending boundaries of returned segments. Phonetic and partial matching was utilized to correct the errors from automated speech recognition and noun phrase chunking. Further, extra knowledge such as lecture slides were used to facilitate the ASR transcript error correction. The approach also makes use of proximity to approximate the deep parsing and structure match between question and sentences in ASR transcripts. The experiment results showed that both phonetic & partial matching improved the segmentation performance; proximity was an effective approach to improve the overall performance.

The chapter is organized as follows. Section 5.1 introduces the motivation of the study. Section 5.2 reviews the related literature. Section 5.3 presents the design and implementation of the query specific segmentation approach. Section 5.4 discusses the experiment and findings. Section 5.5 concludes the chapter by summarizing the contribution and pointing out the future directions.

5.1. Motivation of the Study

While Chapter 4 and 5 focus on solving the problem of extracting the content structure of a lecture video (the static topic segmentation), this chapter addresses the challenge of retrieving relevant segments based on user query. Previous research from LBA (Zhang and Nunamaker 2002) addresses this problem partially by providing search capability to allow students to ask a question with keywords or natural language question on a corpus of lecture video. The LBA system will return a list of video clips as answers. However, these video clips are pre-segmented with fixed sizes and hard to fulfill a large variety of users' information need. The system will not work properly if the users ask for a segment smaller or larger than the fixed size segment. For example, for a question such as "What is the most popular search engine", an one sentence video segment (e.g. 10 second) like "Google is the most popular search engine …" is a better answer than a pre segmented 10 minutes video clip although it does contains the answer. With the redundant information and the answer hiding in the 10 min video, a significant amount of time is still required to find the exact answer. A dynamic segmentation approach that can return arbitrary sizes of segments according to user queries provides more accurate answers and saves human labor.

Different from search engines such as Google (http://www.google.com/), a dynamic segmentation approach returns a chunk of information instead of a full document, which is similar to passage retrieval in the information retrieval area. However, the dynamic segmentation approach is unlike most passage retrieval methods because the sizes of

returned segments are determined on the demand of a user's query (we use "dynamic segmentation" or "query-specific segmentation" interchangeably for the rest of the chapter). The dynamic segmentation approach also varies from the methods explored by text question answering community (Voorhes 2000). The approach works on a broader and more complex category of question types and videos (specifically lecture type videos) instead of text. The uniqueness of the research problem is summarized after reviewing the related research.

5.2. Related Research

5.2.1. Question Answering

Automated question answering is the technology that allows a user to ask a question using everyday language (natural language questions) and receive answers quickly and succinctly. The TREC question answering track (http://trec.nist.gov/data/qa.html) is the motivation force for recent successes in question answering research. While the subject matter of the questions is not restricted, the question types are mainly restricted to fact-based, short answer questions (e.g. *how many calories are there in a big Mac*?). Automated Question Answering (QA) relies on many technologies, mainly Information Retrieval (IR) and Natural Language Processing (NLP). Most QA systems use a variant of the following three steps (Voorhes 2000).

1. Question Understanding. The system first attempts to recognize and classify question type. For example, a question with “who” (e.g. who is the CEO of IBM?) implies the answer to be a person.
2. Document Filtering. Next the system retrieves a small number of documents using standard document retrieval technology and keywords in the question (with expansion such as morphological or using WordNet).
3. Answer Extraction. Finally the returned small portions of documents are parsed to detect the same type of entities as the answer (e.g. “Lou Gerstner”).

While some TREC participants used shallow NLP and pattern matching (Rabagiu et al 2000, Soubbotin and Soubbotin 2002, Wu et al 2003), others relied on deep NLP techniques (Moldovan et al 2003, Scott and Gaizauskas 2000). The most successful system within the last few years, Power Answer (Falcon) (Moldovan et al 2003), relies on a pre-built hierarchy of dozens of semantic types of expected answers, complete syntactic parsing of potential answer sources, and automated theorem-proving to identify the answers. The system first parses the question into a semantic form and uses it to determine the expected answer type by finding the most connected phrase; then the system retrieves paragraphs from a corpus using different combinations of terms in the question and its expansion (e.g. from WordNet); the retrieved paragraphs are parsed into semantic forms; unification and logic proving are performed between the question and paragraph semantic forms to retrieve the answers. It also includes successive feedback loops that try larger modification of query terms to the original question until it finds an answer that can be justified as an abductive proof.

In this chapter we explore how text-based QA techniques can be applied to videos, more specifically lecture videos. Although video retrieval has been studied for decades, question answering on videos is a relatively new research area. Existing video retrieval systems return either a whole video sequence or a pre-defined summary in response to a user query. As discussed in Chapter 1, it is often difficult and time consuming for people to find specific information in video streams. On the other side, existing techniques used in text-based question answering can not be simply applied to videos. Retrieving concise and informative answers from videos requires a good understanding of the video semantic content. The semantic content often comes from multiple sources including video images, audio, accompanying speech or closed captions, and metadata. Beside the requirements of fusing multiple sources of content information, these sources of content may contain errors or be inconsistent (e.g. speech recognition errors in speech transcript). These errors need to be corrected before a question answering system on videos can achieve a good performance.

5.2.1.1. Video Based Question Answering

Yang (2003) presented a question answering system on news videos called VideoQA. The VideoQA system uses short natural language questions with implicit constraints on contents, context, duration, and genre of expected videos. The system returns a summary of relevant video segments as the answers, supplemented by text versions of latest news. The general question answering process is described as follows.

- At first, each video is segmented into two levels: shot level (each shot is classified into a genre such as weather, or sports) and story level.
- Then words, noun phrase, name entities, and answer target are extracted from the question. The question is further reinforced by extracting context related words from recent news articles relevant to the video.
- Finally, the extracted components in the expanded question (noun phrases, name entities, query words, expanded words, answer target, video genre) are matched with each sentence in the speech transcript. The final matching score for each sentence is:

$$S_{ij} = \sum_{k} \alpha_k W_{ik} \tag{5.1}$$

Where α is the weight, and W is the similarity score for each component. Top ranked sentences above certain threshold are returned as answers. To overcome the speech recognition errors, name entities extracted from both recent news articles and video text (from OCR output) are used to match phonetic similar name entities in the transcript.

As the above VideoQA research focuses on the genre of news videos, Zhang (2002) and Cao et al (2004, 2005) explored the research of question answering on lecture videos. Zhang (2002) applied a natural language-based QA approach to a collection of transcribed lecture video documents. In that approach, the transcripts are segmented manually into topics during preprocessing. A three-step approach similar to those in text-based QA is used. However, unlike most text-based QA systems, it uses a template-based

approach for question understanding and answer extraction. The approach uses a parser called Conexor iSkim (Voutilainen 2000) to extract major verbs, nouns, noun phrases, named entities. Their synonyms are also extracted using the WordNet dictionary. A question is filled into a question template with 9 slots: answer type, question focus, person, organization, governor, objects, number, time, and location. Each sentence in the speech transcript is parsed and transformed into sentence templates (ST) in a similar form as the question template (QT). Similarity between the QT and the ST is calculated based on the 9 slots of QT and ST. Finally, a sliding-window method is used to calculate the total similarity between the question and each five-sentence window in each pre-divided segment of the speech transcript. The highest score is taken as the relevancy score of the segment and the top relevant segments are returned. Cao et al (2004, 2005) used a similar template matching approach. Furthermore, the text extracted from PowerPoint slides associated with the lecture videos is used as a source of domain knowledge to boost the answer extraction performance. PowerPoint slides are also used to correct the speech recognition errors in the transcript. The research found that the PowerPoint slides improve the QA performance. The approach works best for human generated transcripts plus the help from PowerPoint slides.

Although the question answering systems from Zhang (2002) and Cao et al (2004, 2005) are designed for lecture videos, both systems require human generated speech transcripts. Human generated transcripts are not available for most lecture videos and generating speech transcripts manually is very time consuming. Furthermore, both approaches

assume that the videos have been pre-segmented into topics by human. As discussed before, manual segmentation is very time consuming and pre-divided video segments can not fulfill the users' various information needs.

5.2.2. Passage Retrieval

Compared with document retrieval, passage retrieval has several advantages. It provides shorter and convenient units of text to users, avoids the difficulties of comparing documents of different length, and enables the identification of short blocks of relevant material amongst otherwise irrelevant text (Kaszkiel 1997). Retrieval based on passages is either based on division of documents according to document markups such as sections (Salton et al 1993, Wllkinson 1994), paragraphs (Zobel 1995), fixed-length sequences of words (Callan 1994), or boundaries given by inferred shift of topic (Hearst and Plaunt 1993, Knaus et al 1995, Mittendorf and Schauble 1994). However, all the above research works involve an additional and preliminary stage of extracting passages from documents, and there is very little research on extracting passages dynamically. One exception is Mittendorf and Schauble's work (1994), which infers passage boundaries by employing a Hidden Markov Model (HMM) to determine passages appropriate to each query. The proposed approach assumes that a document was produced by a HMM and there are certain states in the HMM which model the production of a passage relevant to a query and other relevant for producing non-relevant passages. The Viterbi algorithm is used to find the most probable state sequence.

5.2.2.1. Passage Retrieval in Question Answering

Recently, passage retrieval has become an important component common to many question answering systems. Many current question answering systems have four components/steps instead of three: question analysis, document retrieval, passage retrieval, and answer extraction (Tellex et al 2003). After the document retrieval component finds the relevant documents, the passage retrieval component usually selects a handful of paragraph-sized segments. The simplest reasonable passage retrieval algorithm counts the number of terms a passage has in common with the query, where each sentence is treated as a separate passage (Light et al 2001). The MultiText algorithm (Clarke 2000a, 2000b) identifies short text fragments or "hotspots" where query terms appear in close proximity. The score of a hotspot is computed from its length and the weights of the terms it contains. IBM's passage retrieval algorithm (Ittycheriah 2001) computes a series of distance measures for the passage by summing up the IDF values of words appearing in both the query and the passage. SiteQ's algorithm (Lee et al 2001) computes the score of an n-sentence passage by summing the weights of the individual sentences. Sentences are weighted based on query term density and their part of speech. ISI's passage retrieval algorithm (Hovy et al 2001) ranks sentences based on their similarities to the question by weighing various features: exact match of proper names, match of query terms, and match of stemmed words.

As indicated in a recent quantitative evaluation of passage retrieval in question answering (Tellex et al 2003), common to all best performing algorithms in the evaluation is a

density-based weighting of query terms (e.g. they favor query terms that appear close together). The authors believe that density-based scoring is a critical aspect of passage retrieval. Besides being an important intermediary between full documents and exact answers, passages themselves form a very natural unit of response for question answering systems. Lin et al (2003) showed that users preferred passages over exact phrase answers in a real-world setting because paragraph-sized segments provide contexts of the answers.

5.3. Research Objectives and Challenges

As described at the beginning of the chapter, our research objective is to design and develop a dynamic segmentation algorithm that accepts user queries by natural language questions, and returns just the specific chunk of information that users need. The research is similar to question answering (including video-based question answering) and passage retrieval research, but different in the following ways:

- *Complex Question Types*: Similar to text-based and video-based question answering research, dynamic segmentation accepts questions instead of keywords. While most question answering research focuses on factoid questions, our research intends to answer more complex question types: mainly definition, list, and procedure questions. According to our observations from the log files, these three types of questions were mostly asked in the question answering-based learning systems LBA (Zhang, 2002), and its extensions, Agent99 Trainer (Cao et al 2004). One common feature of the three question types are that they usually ask for one or more fragments of information rather than phrases as in factoid questions. For example, the question

"*How do we define nonverbal communication*" looks for a fragment which contains the definition of nonverbal communication, with or without further explanation. However, compared with factoid questions these questions (especially *how* and *what* questions) tend to be more difficult to answer because of their insufficiently narrowed answer types. For example, *what* questions provide little constraint on the answer type (e.g. *what is visual primacy* vs. *what's the impact of nonverbal communication in human interaction*).

- *Dynamic Segment Sizes*: dynamic segmentation has the same goal as passage retrieval component a in text-based question answering system in terms of returning passages instead of full documents. However, first our dynamic segmentation approach needs to work on videos instead of text. Second, unlike the text domain the existing video-based question answering systems do not incorporate a passage retrieval component, or simply treat a sentence as a passage (Yang et al 2003). Finally, most passage retrieval techniques used in question answering literature are described in the context of improving the answer analysis performance. The returned passages are in a fixed size. On the contrary, we are looking for a method to determine the sizes of the answers dynamically, according to a specific question asked by the user. However, several challenges remain to develop methods to return answer segments with dynamic sizes: 1) First, it is often difficult to decide the sizes of the answer segments. Should it be smaller to be more concise, or larger to include more contexts? Without interactions with users, it is not easy to determinate how much information the user is seeking. 2) Second, it is hard to find the locations where the answer segment should

start or end. The words, phrases, or patterns may do not happen in certain video segment although the segment is a valid part of the answer.

- *Lecture Video Characteristics*: First of all, question answering on videos are inherently more difficult than question answering on written text because 1) the content structure of a video is often hidden in multiple sources (images, speech, audio, and metadata), and their space relationships and temporal timeline; 2) Errors and inconsistency happen frequently because of the limitation of existing technologies (e.g. speech recognition errors). Second, unlike films and news videos, lecture videos' special characteristics (Chapter 2) introduce more challenges to the dynamic segmentation process: 1) Lecture videos are usually produced in an unprofessional or informal way compared to professionally and commercially produced films or news videos. Speech accompanied with video is usually unscripted, spontaneous, and lacks structure. Transcript quality is expected to be worse and human generated transcripts such as closed captions (in news video) are usually unavailable; 2) Knowledge sources such as recent news articles used in VideoQA system (Yang et al 2003) do not exist for lecture videos. However, on the other hand, other forms of extra knowledge sources such as electronic slides may exist and can facilitate the dynamic segmentation. Finally, although lecture videos are produced in a less professional manner, heuristics can still be explored based on camera motions, general instructional methods, or presentation styles especially for a specific instructor or course. For instance, most instructors tend to introduce a concept first before explaining it in order to make the lecture clear and well structured.

The following sections discuss the design and evaluation of such a dynamic segmentation approach, called *DynamicSeg*, to address the above challenges. Users interact with *DynamicSeg* by asking a natural language question. *DynamicSeg* will search a corpus of lecture videos, find the most appropriate video segments, and return to users.

5.4. The Approach

5.4.1. System Architecture and Overview

DynamicSeg aims to provide precise video segments as answers to complex question types such as definition questions and procedure questions. Figure 5.1 describes the system architecture of *DynamicSeg*.

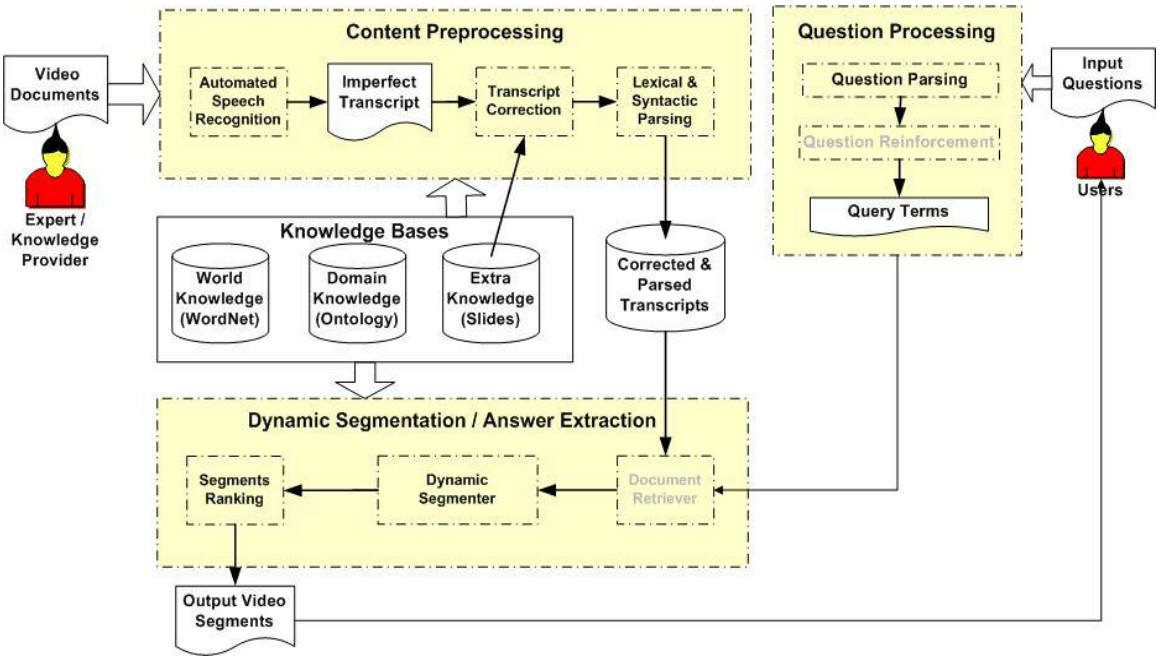


Figure 5.1. System architecture for *DynamicSeg*

Given a lecture video corpus, the content preprocessing stage prepares the videos for later retrieval. At first, we use an automated speech recognition engine called Virage VideoLogger (http://www.virage.com) to extract the speech transcripts from the audio tracks of the videos. Then a phonetic matching-based transcript correction method is applied to correct the speech recognition errors. The transcript correction method focuses on correct words (especially noun phrases) wrongly recognized as one or more words with similar sounds. Finally, the corrected transcripts are parsed and terms such as noun phrases and word stems are extracted and stored for later retrieval purpose.

The question answering stage includes two steps: *question analysis* and *dynamic segmentation*. A user's question is first processed in the *question analysis step*. Terms such as noun phrases and word stems are extracted. In order to match words wrongly recognized in the transcripts, the phonetic forms of noun phrases in the question are also generated and used to match phrases with similar sounds in the transcripts. We leave more complicate question analysis to future research because difficulties to determine the answer types for *how* and *what* questions, which happen most frequently in our question types. For instance, while "*what is visual primacy*" is a definition question, "*what countries tend to express emotions more openly"* is looking for locations, the question "*what are the differences in the facial appearance between least and best attractive faces in females*" may require an answer with multiple segments from multiple documents. The second and most important reason why we leave the complex question analysis to future research is because we want to focus on the segmentation part. We are more interested in

how the segmentation only (without complex question analysis) can improve the question answering performance.

In the *dynamic segmentation* step, at first we need to retrieve the top ranked documents. Because our research focus is on the segmentation part, we use an oracle document retriever for simplification purposes. The oracle document retriever always returns the documents which contain the correct answer. We implement that by hard coding the document ID for each question, similar to the method used in (Tellex et al 2003). Then we use a sliding window method to scan the document and find the best chunks of answers. Phonetic matching and partial matching has also been used to correct two types of errors: wrongly recognized words due to speech recognition errors and wrongly recognized noun phrases due to poor performance of noun phrase chunker. The details of *DynamicSeg* are described as follows.

5.4.2. Preprocessing

In the preprocessing stage, the text transcripts of lecture videos are generated by a speaker-independent speech recognition tool: Virage VideoLogger. The VideoLogger also generates time stamps for each word that synchronize the video stream with the transcript. However, this tool does not generate sentence boundaries and the transcript contains non-negligible speech recognition errors. The transcript is segmented into sentences using a simple heuristic, corrected using text from electronic slides, and then parsed into different feature vectors. Sentence segmentation and transcription correction

are illustrated in details as follows. Parsing will be discussed in the section describing the *dynamic segmentation* step.

5.4.2.1. Sentence Segmentation

Finding sentence boundaries is a necessary first step for parsing in the later steps, including part-of-speech tagging and noun phrase extraction. It is also important because returning complete sentences as answers is more meaningful and easy to understand for end users. Speech sentence segmentation is challenging because the cues typically presented in segmenting written text (headers, paragraphs, punctuation) are absent in spoken language. Currently, we use one simple prosodic feature, pause length, to find the sentence boundaries. Previous research has shown that longer pause durations imply a higher probability of a sentence boundary and pause durations are relatively reliable features (Tur et al 2001).

5.4.2.2 Transcript Correction

Because of the complexity of human vocal tracks, speakers' differences, dialects, transmission distortions, and speaking environments, many errors occur during speech recognition (Yang et al 2003). Compared to speeches in news videos, speeches in lecture videos are more casual and spontaneous, which introduces more potential recognition errors, reported as from 35-60% word error rate (Ponceleon and Srinivasan 2001). Several studies (Hauptmann et al1998, Johnson et al 1999, Thong et al 2002) have shown that typical spoken document retrieval (SDR) systems could tolerate up to 30% speech

recognition errors with no significant reduction in retrieval accuracy. However, these research results are not applicable to our dynamic segmentation task because we are dealing with sentence-level answers rather than document-level answers in SDR. For example, many important terms are often repeated many times, at the document-level in SDR. Thus, it is more likely that they can be correctly recognized at least once. However, this is not the case for the dynamic segment task because most terms often occurs only once in a sentence-level.

Similar to the method used in VideoQA (Yang et al 2003), we focus on one major type of speech recognizer error: substitution. Substitution is when one or more wrong similar sounding words are "substituted" in place of the correct word. Examples are: "Dr. David Buller vs. Dr. David Bowler"; "physical appearance" vs. "physical parents"; and "knowledge test" vs. "knowledge passed". The substitution type of error happens most frequently on name entities (e.g. the name of a person) and noun phrases. To simplify our discussion, we use noun phrases (NP) to denote both noun phrases and name entities since name entities are generally one type of noun phrases. Moreover, we focus on noun phrases because the names of concepts and theories that are essential in a lecture are usually noun phrases. Therefore, the recognition errors of noun phrases must be corrected in order to improve the performance of dynamic segmentation and final question answering.

Previous research works convert words to phonetic sounds and use the phonetic sequence to expand (Chen et al 2001) or correct (Wang et al 2003) spoken language queries for effective information retrieval. We adopted an approach from VideoQA which is based on phonetic matching. Although knowledge sources such as news articles do not exist for lecture videos, fortunately lecture videos usually have accompanying slides (e.g. PowerPoint presentation slides). Slides are a good source of knowledge for lecture videos. As an outline or summarization of a lecture, slides share many common terms with the speech transcript. While the words or noun phrases extracted from slides constrain the list of terms to be corrected, there are still a large number of words with similar sounds. For instance, the phonetic matching algorithm may tend to wrongly correct "intimidating behaviors" with "immediacy behaviors" because they have similar soundings. To avoid mis-corrections and improve precision, we further restrict the correction and phonetic matching to specific text areas in the transcript, which correspond to an individual slide. The details of the transcript correction approach are described as follows.

The first step of transcript correction is to extract potential matching targets from slides, mainly NPs (including name entities). At first, slide texts are extracted from. Name entities (e.g. locations, person, and time) and objects (other types of noun phrases) are extracted using a parser from previous research (Zhang, 2002). Each line in the slide text is treated as a sentence, the unit for parsing. However, the parser often fails because many lines in slides are not complete sentences, but phrases or clauses (e.g. "Communication

Competence", "Focus on the Middle East and Asia"). We use a heuristic to solve this problem similar to the method used in (Syeda-Mahmood and Srinivasan 2000). We extract all possible two or three words phrases from each short line (e.g. less than 6 words). The final result is a list of NPs for each slide (NP_{ij} ; i is the slide number)

The next step is to restrict the transcript correction matching to corresponding transcript segment for each slide. This process constrains the context of where the potential NP occurs, and thus improves the accuracy of correction. We reuse our dynamic segmentation approach (its details will be described in next section) to find the best matched segments. In other words, we treat the slides text as a query; then use the dynamic segmentation approach to find the best segments for the slide text query where $p(B_i \mid S_i) > \delta_1$. B_i is the best segments or blocks with $B_i = (w_{i1}, w_{i2}, \ldots, w_{ip})$ of p terms; S_i is the slide i.

The final step of transcript correction is to find phonetic matching between extracted NPs (NP_{ij}) and B_i . We look for the maximum matching terms which are above a certain threshold δ_2 . The phonetic matching measure is based on the approach from VideoQA (Yang et al 2003). We first extract the phonetic representations of both NP_{ij} and a potential words string in B_i . The phonetic representations are generated using the CMU pronouncing dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). For instance, the phonetic representation of "physical appearance" and its recognition error ("physical

parents") are < F IH Z IH K AH L>, < AH P IH R AH N S> and < F IH Z IH K AH L>, < P EH R AH N T S> respectively.

Given two phonetic strings *x* and *y* approximately the same length, the overall similarity (Yang 2003) is:

$$S_p(x, y) = a_b \, S_b(x, y) + a_l \, S_l(x, y) \tag{5.2}$$

Where $a_b = a_l = 0.5$; $S_b(x, y)$ and $S_l(x, y)$ are defined as follows.

- $S_b(x, y)$ is *String Boundary Similarity*, measuring the similarity in the start (start()) and ending (end()) phonetic sounds. If start(x)=start(y) and end(x)=end(y), $Sb(x, y) = 1$; if start(x)=start(y) xor end(x)=end(y), $Sb(x, y) = 0.5$; otherwise $Sb(x, y) = 0$.
- $S_l(x, y)$ is *Longest Common Sub-sequence* (LCS) Similarity, which computes the number of phonetic matches between two strings in their occurrence order [9], and normalizes the measure by the phonetic length of string x.

  $$S_l(x,y) = LCS(x,y) / |x| \tag{5.3}$$

If NP contains many words (e.g. "physical appearance"), we calculate the overall $S_p(x, y)$ as follows.

$$S(x, y) = \frac{(\gamma)^{N-1}}{N} \sum_{k}^{N} S_k(x_k, y_k) \tag{5.4}$$

One example of the matching between NP and recognized word string is in Figure 5.2.

Original NP: physical appearance < F IH Z IH K AH L>, < AH P IH R AH N S>
Recognized word string: physical parents < F IH Z IH K AH L>, < P EH R AH N T S>
For physical: S_b(x, y) = 1; S_l(x, y) =1 ;
For appearance: S_b(x, y) = 0.5; S_l(x, y) = 5/7;
S_p("physical appearance") = [S_p("physical") + S_p("appearance")]*γ/2

Figure 5.2. Examples of phonetic matching between NP and recognized word string

5.4.3. Dynamic Segmentation

The major task of the dynamic segmentation is to find the start and ending points of the answer segments. We use a sliding window approach which is similar to what is used in our static segmentation (Chapter 4). First, we extract feature vectors (e.g. noun phrases, verbs, word stems) from both question and each sentence in the transcript. We then move a sliding window across the transcript, compute the similarity between the question and sliding window at each position. We identify the start boundaries as locations where the whole sliding window is within the answer segment area, and the ending boundary as locations where the whole sliding window is out of the answer segment area. The details of the approach are described in the following sections.

5.4.3.1. Features Extraction and Syntactic Parsing

Before feature extraction, we preprocess and parse the transcript using a natural language tool called GATE (Cunningham, 2000). GATE handles tokenization and part-of-speech (POS) tagging. The POS tagger (Hepple, 2000) in GATE is a modified version of the

Brill tagger, which produces a part-of-speech tag (e.g. NN for nouns and VB for verbs) as an annotation on each word or symbol. Porter's stemmer (Porter, 1980) is used for suffix stripping. Uninformative words are removed using a stopword list.

We extract five types of feature vectors: noun phrases (NP), verb classes (VC), word stems (WS), partial noun phrases (PNP), and phonetic matching noun phrases (PMNP). Noun phrases (NP) are the joint set of BaseNPs and Nouns (NN). BaseNPs (BNP) are extracted using a noun phrase chunker within GATE. This chunker is a Java implementation of the Ramshaw and Marcaus BaseNP chunker (Ramshaw and Marcaus, 1995). Nouns (NN) are single nouns recognized by the POS tagger. Verb classes (VC) are verb categories classified based on WordNet (Chapter 4). Word stems (WS) are other content words beside NP and VC. We believe shallow syntactic parsing (e.g. POS tagging and noun phrase chunking) will result in better performance than lexical information although many recognition errors occur in the speech transcript. Figure 5.3 gives an example of the syntactic parsing. The POS tags show that the POS tagging performance is pretty good despite the speech recognition errors. Although the words "nonverbal" and "communication" are wrongly recognized as "verbal" and "Haitians", their POS tags are still correct. "Introduce" is recognized as "introduced", but the POS tags are still close to each other (VB vs. VBN).

Partial noun phrases (PNP) and phonetic matching noun phrases (PMNP) are two special features we introduced to resolve the possible failures of matching during segmentation

because of the speech recognition errors. Phonetic matching is used for situation when a query term from the question is wrongly recognized as similar sounding term in the transcripts. For instance, the word “pleasantness” in the question “What are the vocal emotional expressions for pleasantness?” (Q34) is recognized as “was at nets” in the transcript. We may not be able to found the answer because there is no word match between “pleasantness” and “was at nets”. A phonetic matching for similar sounds like the one in transcript correction will solve this problem. We adopt the algorithm of phonetic matching from our transcript correction component (Section 5.4.2.2). We first extract the phonetic representations of the noun phrases in current question. The phonetic representation is compared to each sentence in the transcript using the phonetic matching algorithm. If the similarity is above certain threshold (e.g. δ_3 =0.8), the noun phrase is stored in the feature vector of PMNP.

From our observation, although the performance of POS tagging is reasonably good, we found that the noun phrase chunker made many mistakes in the process of chunking. For example, in Figure 5.3, “introduced few major components” is chunked as one baseNP because “introduce” is wrongly recognized as “introduced” (its past participle form). However, the recognized noun phrase “introduced few major components” does match the correct noun phrase “major components” partially. Thus, partial matching is introduced to handle the special situation where the noun phrase chunker fails, but the wrongly chunked NP still partially matches the correct NP. We calculate the normalized LCS value, $Sl(x,y)$, between correct NP (x) and chunked NP (y) according to formula (2).

Human corrected sentence: This 14-unit series is designed to introduce you to the major components of the nonverbal communication .

Sentence recognized in ASR transcript:
his fourteen units series is designed to introduced few major components on verbal Haitians .

POS tags:
[PRP$, CD, NNS, NN, VBZ, VBN, TO, VBN, JJ, JJ, NNS, IN, JJ, NNS, .]

Noun phrases (after stemming):
[hi fourteen unit seri, introduc few major compon, verbal haitian]

Figure 5.3. An example of the syntactic parsing result of speech transcript

If $Sl(x,y)$ is larger than certain threshold (δ_3 =0.3). Two NPs roughly match, and we return the $Sl(x,y)$ value as matching weight in the Partial Noun Phrase (PNP) feature vector.

5.4.3.2. Sliding Window Approach

The basic idea of the sliding window approach is similar to *PowerSeg* (Chapter 4). We move a sliding window (e.g. 4 sentences) across the transcript by a certain interval (e.g. 1 sentence), also called gap. For each gap, instead of comparing one similarity between two neighboring windows as in *PowerSeg* (we call them left and right windows), we calculate two types of similarities: the similarity between the question and the left sliding window ($Sim(q, w_L)$), and the similarity difference between question, left, and right windows ($Sim(q, w_L, w_R)$). Then we draw two similarity graphs for both $Sim(q, w_L)$ and

$Sim(q, w_L, w_R)$ during all gaps (see Figure 5.4). The gaps with largest $Sim(q, w_L, w_R)$ changes and maximum / minimum $Sim(q, w_L)$ values are the locations for start/ending boundaries. We will describe the details after presenting the similarity measure as follows.

5.4.3.2.1. Similarity Measures

The similarity between question and left window w_L: $Sim(q, w_L)$ is calculated by the cosine measure: the weighted sum of the cosine products of five feature vectors we presented before.

$$Sim_k(q, w_L) = \sum_j \frac{\sum_i f_{j,i,q} f_{j,i,w_L}}{\sqrt{\sum_i f^2{}_{j,i,q} \sum_i f^2{}_{j,i,w_L}}} S_j P_k \qquad (5.5)$$

$f_{j,i,q}$ or f_{j,i,w_L} is the i-th term weight value of j-th type feature vector in the question or left window (j = 1..5, are different feature vectors; and i = 1…m, are different terms in a specific feature vector. $f_{j,i,q} f_{j,i,w_L}$ is calculated based on term frequency (TF), the repetition times of terms (e.g. noun phrases). s_j is the significant value (or weight) for each specific feature vector. Currently s_j values are calculate based on human heuristics and hand tuning. P_k is the what we called proximity value, computed by the following formula.

$$P_k = 1/\sqrt{k} \tag{5.6}$$

k is the number of sentences in the current sliding window (e.g. $k = 4$). The proximity value is a concept similar as the density-based weighting of query terms in passage retrieval (Tellex et al 2003). In other words, it favors query terms that appear close together. In our approach, when calculating the similarity between question and sliding window (e.g. window size = 4 sentences), we do not fix the window size. Instead, we allow window size to vary from 1 to 4. Then we compute the similarity between question and each varied size window and use the maximal value as the final similarity values between question and sliding window (w_L here).

$$Sim_k(q, w_L) = \arg\max_k Sim_k(q, w_L) \tag{5.7}$$

$Sim_k(q, w_L)$ is calculated based on (4) for each window size (e.g. $k = 1 \ldots 4$) . It is easy to notice that if we only restrict the window to consecutive sentences, we have $n = \sum_{i=1}^{k} i$ number of different sizes of windows. It will largely increase the computational cost if the window size is relatively large (e.g. $k = 6$, $n = 21$). Different from other density-based weighting schema in passage retrieval literature (Tellex et al 2003), we restrict the windows to only those that start from sentence 1. For instance, if $k = 4$, we only have 4 windows (sentence 1; sentence 1 to 2, sentence 1 to 3; and sentence 1 to 4). The heuristic here is that we assume that the first few sentences are more important than others because most instructors or speakers will clearly introduce the concepts, theories or topics before explaining them. Our observation on several datasets confirms the heuristics. People

intend to explicitly outline or define a concept or topic first when they are teaching or presenting. It also agrees with our common sense because instructors are taught to teach in such a manner.

Calculating the similarity between question, left window (w_L), and right window (w_R) is not an easy task. This similarity value tries to capture the trend of the similarity between question and sliding window when the window is moving across the transcript. Currently we use a simple way to calculate this value: the difference between $Sim(q, w_L)$ and $Sim(q, w_R)$.

$$Sim(q, w_L, w_R) = Sim(q, w_L) - Sim(q, w_R) \quad (5.8)$$

$Sim(q, w_R)$ is calculated based on the same formula (5.4) as $Sim(q, w_L)$.

5.4.3.2.2 Identifying the Start and Ending Boundaries

To identify the start and ending boundaries of the answer segment, we move a sliding window (e.g. size = 4 sentences) across the transcript by certain interval (e.g. 1 sentence) (the top part of Figure 5.4: window sliding). Let's assume that the correct answer segment size is smaller than the size of the sliding window (e.g. the correct answer is from sentence 199 to 202, excluding 202) (Figure 5.4). The situation when the answer size is larger than the window size will be similar. Let's further assume that other parts of the transcript excluding the answer segment are irrelevant (or insignificantly relevant) to

the question. Roughly, the similarity between query and sliding window depends only on the size of the overlapped area.

We move the sliding windows (e.g. left window starting from 195, right window from 1999) across the transcript (Figure 5.4). When the sliding windows keep moving right, $Sim(q, w_L, w_R)$ is becoming smaller because w_L starts to overlap with the answer segment (from no overlap) and w_R has less overlap with answer. The similarity difference between two neighboring windows w_L and w_R ($Sim(q, w_L, w_R)$) reaches the *last lowest point* when w_L's right side and w_R's left side meet the answer's left boundary (Figure 5.4). At the same location, $Sim(q, w_R)$ reaches its first highest value. That location (sentence number 199 in Figure 5.4) is where we should draw a line for the start boundary. Similarly, the ending boundary is at the location where w_L's right side and w_R's left side meet the answer's right boundary. Then, $Sim(q, w_L, w_R)$ reaches its first highest value and $Sim(q, w_R)$ arrives at its first lowest value. In summary, the rules to identify boundaries are as follow.

1. The start boundary is identified where $Sim(q, w_L, w_R)$ reaches its last lowest value and $Sim(q, w_R)$ arrives at its first highest value.
2. The start boundary is identified where $Sim(q, w_L, w_R)$ reaches its first lowest value and $Sim(q, w_R)$ arrives at its first lowest value.

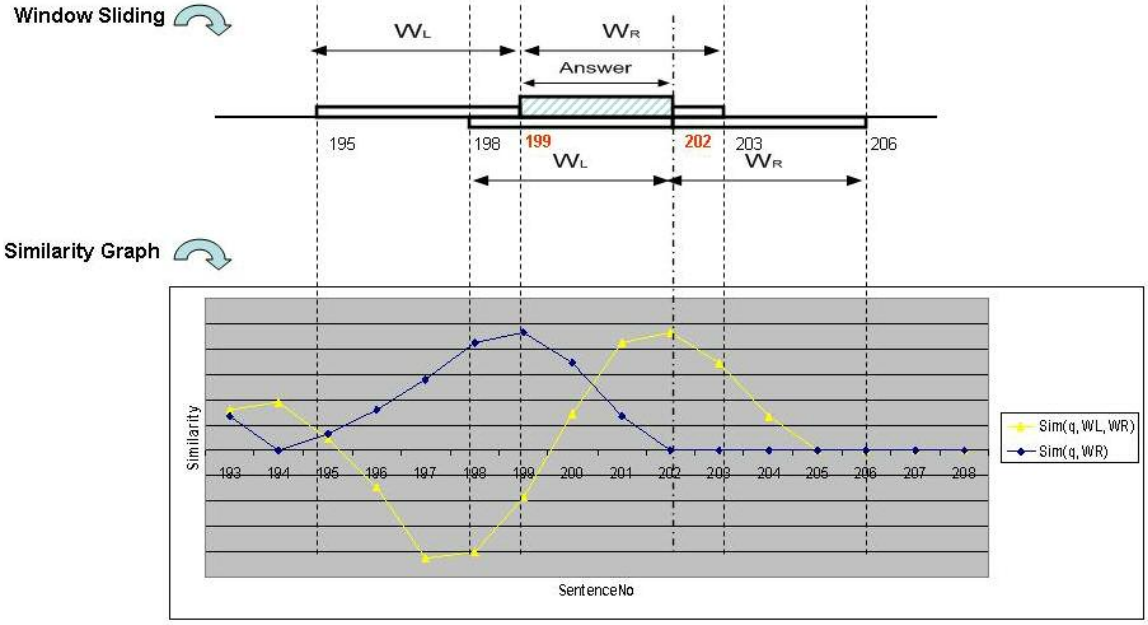


Figure 5.4. Identify boundaries: A sliding window approach

The bottom part of Figure 5.4 shows an example how the start and ending boundaries are identified using the rules described above. The similarity graph was draw based on the question "*How do information, behavior, and communication related to one another?*", and a speech transcript from a nonverbal communication course (section 5.5). The yellow line is $Sim\ (q, w_L, w_R)$, and the blue line is $Sim(q, w_R)$. We notice that the ending boundary (sentence 202) corresponds to the highest value of $Sim\ (q, w_L, w_R)$ and first lowest value of $Sim(q, w_R)$ perfectly. However, the lowest value of $Sim\ (q, w_L, w_R)$ and first highest value of $Sim(q, w_R)$ are not at the same location. It is because the assumption that other parts of the transcript excluding the answer segment are irrelevant to query

does not hold in real situation. Sentences that are not the answer segment could also contain query terms from the question. At this circumstance, we identify the start boundary as the location where $Sim(q, w_R)$ has the highest value (sentence 199). The highest $Sim(q, w_R)$ rule overrides the requirement of lowest $Sim\ (q, w_L, w_R)$. However, it is mandatory that $Sim\ (q, w_L, w_R)$ should have a low value at least less than certain threshold (e.g. *Mean – SD*).

5.5. Evaluation

5.5.1. Dataset, Hypotheses and Measures

To test the effectiveness of our dynamic segmentation algorithm (called *DynamicSeg*) on answering complex questions on lecture videos, we selected 12 videos from a course about non-verbal communication as our testing data set. The purpose of the course is to introduce the major components of the nonverbal communication system and the communication purposes it serves. The videos have various types of narrative elements such as instructor's talking head, slide, and demo. The average length of the videos is around 42 minutes, which it gives a total of 512 minutes of lecture videos (Table 5.1). The average length of the speech transcripts are 5847 words and 314 sentences. The PowerPoint slides used by the instructors are also used as extra knowledge sources for transcript correction.

Table 5.1. Statistics of the dataset

Video/Transcript	NoOfWords	NoOfSentences	Duration
L01	5552	365	42.22
L02	5658	282	41
L04	6971	424	46.29
L05	3715	183	29.25
L07	6894	264	58.24
L08	7345	433	52.58
L09	5552	243	36.17
L10	3430	193	28.45
L11	6163	321	43.57
L12	7932	442	56.54
L13	5741	322	40.11
L14	5212	293	37.34
AVG	5847	314	42.65

We designed 50 questions related to the 12 lecture videos. Among the 50 questions, 10 are definition questions, 14 questions are list type, 19 are procedure/relationship questions, and only 7 are fact based questions. Therefore, most questions belong to the three question types we are most interested in: definition, list, and procedure questions. The 50 questions are listed in Figure 5.5.

Beside the overall performance of our proposed approach, we are also interested in each component's contribution on overall performance. More specifically, we are interested in the following hypotheses:

H1: The introduction of shallow syntactic parsing (noun phrase chunking and POS tagging) improves the algorithm performance.

H2: Phonetic and partial matching improves the algorithm performance.

01 Who are the leading experts in non verbal communication at the lecture series?
02 What are all the reasons for the importance of non verbal communication in daily transactions?
03 How do we define nonverbal communication?
04 What is the impact of nonverbal communication in human interaction?
05 How do information, behavior, and communication related to one another?
06 Explain how non verbal communication played a role in 1960 elections?
07 Give an example where the cultural difference resulted in difference in immediacy levels?
08 What are the types of display rules?
09 What is the definition of communication competence?
10 List the four major components of non verbal skills.
11 What countries tended to express emotions more openly?
12 How do we define physical appearance?
13 What are the characteristics of males with ideal bodies?
14 What are difference in the facial appearance between least and best attractive faces in female?
15 What's the typical funny or jolly somatype?
16 What's "Halo Effect"?
17 Which law enforcement technique relies on physical appearance cues?
18 How do babies between three to six months old communicate?
19 When does the child's ability to decode human intonations increases rapidly?
20 What is the preferred vocal pattern in the United States?
21 What are the qualities of voice?
22 What are the similarities between human and other primates in vocal expression (based on observation from chimpanzees)?
23 What are the characteristics of information transfer in a high context culture?
24 How the brain processes nonverbal and verbal behavior?
25 What is visual primacy?
26 Under what conditions do people place greater reliance on nonverbal than verbal cues?
27 Which factors are used to identify badges?
28 How dress style differentiates culture?
29 Give three reasons that genders differ in nonverbal behaviors?
30 How kinesic behavior differs between men and women?
31 What are the types of culture based on the amount of contact?
32 What is neurocultural theory?
33 What are the four cues most commonly used to express emotion?
34 What are the vocal emotional expressions for pleasantness?
35 How behavior reflecting intimacy, trust, and liking can be identified?
36 How is eye behavior associated with intimacy and attraction?
37 How vocal cues affect the intimacy level of a given interaction?
38 What is monochronic orientation?
39 What are the two kinds of seating arrangement?
40 What are the characteristics of interaction management?
41 What is Interactional synchrony?
42 What cues speakers use to regulate turn taking in conversations?
43 What are the common nonverbal leave taking cues?
44 How do appearance cues affect persuasion and behavioral compliance?
45 How does the use of self adaptors affect persuasive success?
46 How do vocal pleasantness cues of fluency and pitch variety relate to credibility, attraction, and persuasion?
47 How are dominance and status signaled?
48 How accurate people successfully detect deception?
49 What are the reliable kinesic indicators of deception?
50 What is leakage hypothesis?

Figure 5.5. Test questions

H3: The introduction of proximity in similarity calculation improves the algorithm performance.

H4: The algorithm performed on automated corrected transcripts achieves a better performance than the algorithm on original ASR transcripts.

The performance of *DynamicSeg* is evaluated based on two types of measures: answer accuracy and information coverage, which are defined as follows.

- *Answer accuracy* is the percentage of correctly answered questions. An answer is correct if any of the top *m* segments returned by the algorithm matches the answer segment from human. An answer segment from the algorithm is considered to be matched with the human answer if their start and ending boundaries are the exactly same or only a few sentence away (e.g. relax rate = 2). Further, if the human-generated answer contains more than one segment, the algorithm-generated answer is correct depending on different situations:
  1) If the human think that all segments have to be included in order to make an answer correct, then any of the top *m* computer-generated segments have to match all the human generated segments and no less or more than that.
  2) If the human think any single segment of the *n* human-generated segments is a good answer for the question, then the algorithm generated answer is correct if any of the top *m* algorithm-generated answers matches any of the *n* human answer segments.

- *Information coverage* is measured by precision, recall, and F-Measure. The overall information coverage is the average information coverage from all the questions. For each question, the information coverage is the maximum coverage between all algorithm-generated answers and any correct answer from human, defined as follows.

$$P(recision) = \arg\max_{m,n}\left(\frac{\#\_of\_sentences\_matched_{m,n}}{number\_of\_sentences\_in\_computer\_answer_{m,n}}\right)$$

$$R(ecall) = \arg\max_{m,n}\left(\frac{\#\_of\_sentences\_matched_{m,n}}{number\_of\_sentences\_in\_human\_answer_{m,n}}\right)$$

$$F(measure) = \frac{2PR}{P+R}$$

No_of_sentences_matched is the number of sentences correctly matched between algorithm-generated answer and human-generated answer. *m* is the number of algorithm-generated answers and *n* is the number of all possible correct answers from human for an individual question.

5.5.2. Results and Discussion

At first, we extracted the speech transcripts from all 12 videos and each transcript is segmented into sentences using the pause duration heuristic we discussed before (Section 5.4.2.1). An empirical value of 0.6 second was used for the pause duration used in the sentence segmentation. The heuristic-based sentence segmentation approach achieved a good performance with a Precision of 0.8, and a Recall of 0.78. After we segmented and parsed each sentence in the transcripts, we tested the *DynamicSeg* system using the 12

transcripts and the 50 question listed in Figure 5.5. Each question was submitted to *DynamicSeg* and a list of segments was returned as answers. Each individual segment was a potential answer and was compared to the answers provided by three experts. For each question, an individual answer was provided by each expert and the three answers were merged after discussion by the experts. The two measures (*answer accuracy* and *information coverage*) were calculated for the 50 questions. We used a relax rate of two sentences, which means that boundaries two sentences away from the actual boundaries are still considered correct. Currently, we did not rank the list of segments/answers for each question and the answer is correct if any of the segments matches the expert answer segment. It is reasonable because the average number of answer segments per question is relatively small (around 11). In the future, we can rank the answers and select the top three answers for testing. We tested the overall performance by fine tuning and the four hypotheses by adjusting the weights of the four feature vectors, namely noun phrase (NP), partial NP (PNP), phonetic matching (PMNP), and word stems (WS). The Verb Classes (VC) feature is not used because our preliminary testing showed that it has a very small effect on the overall performance.

5.5.2.1 Overall Performance and Hypotheses Testing

The best performance of *DynamicSeg* and the testing results of the four hypotheses are listed in Table 5.2. The discussions are as follows.

Table 5.2. *DynamicSeg* best performance and hypotheses testing

Testing Items		Answer Accuracy	Information Coverage		
			Precision	Recall	F-Measure
Best Performance		0.52	0.65	0.57	0.61
H1	with Syntactic Parsing	0.52	0.65	0.57	0.61
	w/o Syntactic Parsing	0.52	0.65	0.44	0.52
H2	with Phonetic & Partial Matching	0.5	0.57	0.52	0.54
	w/o Phonetic & Partial Matching	0.26	0.34	0.21	0.26
H3	with Proximity	0.52	0.65	0.57	0.61
	w/o Proximity	0.42	0.58	0.45	0.51
H4	with Transcript Correction	0.50	0.66	0.63	0.64
	w/o Transcript Correction	0.52	0.65	0.57	0.61

With fine tuning on the weights of different features, the *DynamicSeg* achieves a best performance of 0.52 in terms of answer accuracy and 0.61 in terms of F-Measure. However, the best performance is achieved without the NP feature, which is against our hypothesis that noun phrases are most important features and should carry the largest weight among all features. A close examination on the results shows that there are many noun phrase chunking errors, which deteriorates the performance. The poor performance of the noun phrase chunking is due to the speech recognition errors, and the NP chunker we used is designed for written text. However, in the best performance weights setting, partial NP (PNP) carries a weight of 0.6. This proves that PNP helps the error correction introduced by NP chunking and NP is still important feature.

H1 was tested by comparing the running results from *DynamicSeg* with all features (with syntactic parsing) to the results from *DynamicSeg* with word stem (WS) only (w/o syntactic parsing). H1 is not supported in terms of answer accuracy. One explanation is the poor performance of noun phrase chunking largely deteriorates the overall performance. However, *DynamicSeg* with syntactic parsing has a better recall score on information coverage than *DynamicSeg* w/o syntactic parsing although there is no difference between precisions. A possible interpretation is that syntactic parsing did introduce better capability to find possible answer segment, which caused the higher recall. However, because of the poor performance of parsing, it is hard to identify the exact locations of the answer segments, which caused the low answer accuracy and no improvement on information coverage precision.

H2 was tested by comparing the running results from *DynamicSeg* with all NP features (NP, PNP, and PMNP) to the results from *DynamicSeg* with NP only. *DynamicSeg* with partial and phonetic matching (PNP and PMNP) achieves a significant improvement on the performance than *DynamicSeg* without partial and phonetic (answer accuracy: 0.5 vs. 0.26 and information coverage F-Measure: 0.54 vs. 0.26). This proves that partial and phonetic matching did facilitate the correction of errors from speech recognition and NP chunking. However, because we did not run a testing on PNP or PMNP feature separately, there is no way to tell how much contribution each individual feature introduces. We are planning to include that in our future studies.

H3 was tested by running two *DynamicSeg* programs with or without the proximity value P_k . Results show that proximity does improve the performance significantly (answer accuracy: 052 vs. 0.42, and information coverage: 0.61 vs. 0.51). This shows that proximity is a good approximation of the answer structure during the matching process.

H4 was tested by running *DynamicSeg* on two different sets of transcripts (with or without transcript corrections). The results are out of our expectation. H4 was not supported. The performance from corrected transcripts is even worse than that from uncorrected transcripts (0.50 vs. 0.52). After a close examination of the corrected transcripts, we realize that transcript correction algorithm mis-corrected many words, which may cause the decrease of the answer accuracy.

5.6 Contribution and Future Directions

In this chapter we discuss a novel approach (*DynamicSeg*) that dynamically segments a lecture video into relevant sizes of segments, which answer a specific question asked by users. The first novelty of the approach is that it is the first question answering system on lecture videos which returns dynamic size of video segment as answers, according to our knowledge. The system also incorporates unique features such as transcript correction using slides, syntactic parsing on ASR transcript, and partial and phonetic matching. Furthermore, it applies a sliding window approach which has the capability to dynamically identify the start and ending boundaries of answer segments. The system achieved a best answer accuracy of 0.52 in our evaluation, which is very promising. With

the incorporation of further steps of complex question analysis and answer refinement and extraction, the performance is expected to be further improved.

However, there are several limitations of our study. First, although the evaluation results are encouraging, one should note that only videos from one course were used in our experiment. Caution needs to be taken when interpreting our findings. More evaluations on larger sets of data from different instructors and courses are needed to increase the reliability and validity of our results. Second, the system will not be able to find answers if the words in the question do not occur in the transcript although there maybe synonyms. The problem is worse for short questions such as "*what is deception detection*" because the shortage of query terms (only deception detection). Query expansion is needed in order to retrieve more query terms. However, the methods using WordNet (e.g. the method used in LBA) is not applicable because WordNet only provides synonyms for general words, not domain phrases such as "deception detection". To solve this problem one idea is to use the Web as a knowledge source and submit the domain phrases as queries to a search engine (e.g. Google) and extract highly co-occurred words/phrases from top ranked documents returned by the search engine. Third, another weakness of our system is that it does not incorporate question analysis component. Question analysis is extremely helpful for questions type such as definition question. For the question of "*what is deception detection*", using matching patterns such as "deception detection is …" or "deception detection is defined as …" will largely improve the accuracy of the dynamic segmentation and the final question answering.

CHAPTER 6

CONCLUSION AND FUTURE DIRECTIONS

This dissertation investigates the design of automated segmentation algorithms to facilitate information seeking in lecture videos. The segmentation research aims to address the challenges of extracting the topic structure of a lecture video and retrieving relevant video segments based on user queries. Studies were conducted to understand the human segmentation phenomenon and extract potential segmentation features and methods. A set of static and dynamic segmentation algorithms was developed based on results from manual segmentation studies. The rest of the chapter is organized as follows. Section 6.1 summarizes the main conclusions of and contributions of the research. Section 6.2 discusses the future directions. Section 6.3 addresses the practical implication of the research.

6.1. Conclusions and Contributions

The first part of the dissertation explores how humans perform manual segmentation on lecture videos and collecting features and rules for the design of an automated segmentation method. The first study is focusing on understanding the segmentation phenomenon such as the content structure of a lecture video and whether they are reliable features across courses/instructors. The second study narrows down to a specific course

to extract potential segmentation features and methods. Beside the finding of the characteristics of a lecture video content structure (linear structure is preferred and topic boundaries are usually fuzzy), the main findings of the two studies are: 1) a two-phase segmentation process: rough segmentation and segmentation refinement; and 2) a list of potential segmentation features in Table 3.2, and summarized in Table 3.3 with their extraction methods. The major contribution of these findings is that they provide guidelines and a set of potential segmentation features and their extraction methods for scholars (including myself) who are interested in developing an automated segmentation approach for lecture videos.

The second part of this dissertation concentrates on addressing the challenge of extracting the topic structure of a lecture video, or in other words, the development of static segmentation algorithms. The first algorithm tries to capture the overall topic content change (based on overall vocabulary change) by sliding a text window over the speech transcript of lecture video. The algorithm uses natural language processing techniques to extract more salient features (e.g. noun phrases and verbs) than bags of words to improve segmentation performance. Experiment results show that noun phrases are salient features and the algorithm achieves a good performance. The second algorithm is a multimodal method that combines three sources of segmentation features (speech text transcript, audio and video) and makes use of knowledge sources such as world knowledge and domain knowledge. This algorithm also simulates the two phase process (initial segmentation and segmentation refinement) found in manual segmentation.

Experiment results show that the methodology of combining multiple sources of features and two phrase process is promising. The main contributions of the static segmentation research are: 1) two algorithms with reasonable good performance on extracting the topic structures of lecture videos, and 2) a general design framework for multimodal segmentation approaches on lecture videos.

The third part of this dissertation addresses the challenge of retrieving relevant video segment for specific user query. We propose a dynamic segmentation method that can decide the start and ending boundaries of answer segments on demand from user queries. The method accepts natural language questions and uses similarity changes between sliding windows (on ASR transcripts) and user questions to detect the boundaries. Extra knowledge extracted from electronic slides is used to correct speech recognition errors. Phonetic and partial matching is utilized to fix mismatches between query and ASR transcript because of the errors from speech recognition and syntactic parsing. The major contribution is: to our knowledge it is first question answering algorithm on lecture videos which is capable to return dynamic sizes of answer segments.

In summary, the main contributions of this dissertation are two aspects. First, this dissertation provides scholars a set of potential segmentation features, design principles, framework that can facilitate their development of effective automated segmentation approaches on lecture videos. Second, the static and dynamic segmentation algorithms make it possible to develop a practical system for lecture video browsing and retrieval.

Both static segmentation algorithms achieve precision of 77% and a recall of 68%, which we believe is sufficient for practical applications because even human experts do not agree totally with their segmentation results (Hearst, 1994). The dynamic segmentation algorithm achieves an F-Measure of 0.64 for our question set. The proposed static and dynamic segmentation algorithms can be applied to facilitate better browsing, retrieval and full usage of the lecture videos. For example, in the Stanford and LBA and Agent99 e-learning systems discussed earlier in Chapter 1, lecture videos can be segmented automatically using the static segmentation algorithms to support better browsing by students. The dynamic segmentation approach can be extended and used as a question answering system to support better video retrieval for students. These automated segment approaches will save a lot of time and effort which human experts would need in order to manually segment the videos and provide a natural interface (by asking questions) for students to find specific information in lecture videos. Finally, the video segmentation techniques also facilitate the classification of videos into topics. This could allow instructors to share their lectures more easily, e.g., by sharing segments of their lecture videos on certain topics. Further, the features we extracted from speech transcripts and slides such as noun phrases can also be used for indexing, classification, or clustering of videos in e-learning or other video applications.

6.2. Future Directions

The future research of this dissertation consists of three main directions: 1) more and extensive studies on manual segmentation; 2) refinement and improvement on both static

and dynamic segmentation algorithms; 3) empirical studies on how does the segmentation facilitate information searching and learning.

For the first future direction, I plan to conduct more and extensive studies on how human perform segmentation with a large data set of lecture videos across courses, instructors, and majors. We are interested in the comparison of content structure, segmentation features from different styles of lecture videos: slides videos, blackboard videos or mixed presentation videos, and discussion videos. We are also interested in, for the same type of lecture video (slide videos), how the subject matter and instructional style affect the topic structure and segmentation features.

The second main future direction is to further refine and improve the segmentation algorithms we already developed.

- For static segmentation, because there are no reliable features across courses or instructors (except overall content changes), one direction is to employ machine learning method to learn the salient features for a specific course or instructor. However, the machine learning method has to handle the situation of sparse or no training data. One possible approach is to use feedbacks from the instructor or students. First we use the existing algorithms to segment a few lecture videos, then either ask the instructor to correct a couple of them, or deploy them in an existing e-learning system and track students' watching behaviors and use the behavior data to refine the segmentation. For example, if a video segment is too long, there is a large

probability that students will stop the video playing after they realize that rest of the video segment is irrelevant to the current topic. Another possible method on using machine learning with sparse/no training data is to use heuristic or empirical estimation. For instance, we are developing a segmentation approach for slide videos using HMM. The objective of the approach is to segment a slide video into chunks corresponding to slide numbers. The HMM approach treats both video frames with slide image and speech text spoken as instances of underlying topic (an individual slide).

- For dynamic segmentation, we are incorporating question analysis & expansion, and answer refinement & extraction into the existing approach. For query expansion, a general knowledge source such as WordNet is not applicable to lecture videos because most of terms are domain dependent and cannot be found in WordNet. We can use Web as a knowledge source and submit the domain phrases as a query to a search engine (e.g. Google) and extract highly co-occurred from top ranked documents. Other knowledge sources such as slides and electronic textbooks are useful for query expansion too if available. Question analysis is helpful for finding answer patterns. For example, the answer of a question such as" *what is deception detection*" may has a pattern of "deception detection is defined as …" Finally, the boundaries of an answer segment can be further refined by checking whether there are shot cuts or cue phrases close to the potential boundaries.

The third future direction of this dissertation is to conduct empirical studies to investigate how the static and dynamic segmentation approaches facilitate information searching and students' learning in an e-learning system. Our previous study showed that a topic list in e-learning system improved students' learning performance by enabling self-pace learning (Lin et al 2003). However, the topic list used in the system is generated by human experts. We'd like to integrate the automated segmentation algorithms to the e-learning system and are interested in whether the segmentation still improves the learning performance given the segmentation errors from the automated methods. We are also interested in how the dynamic segmentation approach improves the information searching lecture videos and facilitate learning especially for review or assignment purpose.

REFERENCES

Abowd, G. D., Atkeson, C. G., Feinstein, A., Hmelo, C., Kooper, R., Long, S., Sawhney, N. and Tan, M. "Teaching and learning as multimedia authoring: The classroom 2000 project," *Proceedings of Multimedia '96*, pp 187-198, ACM Press, 1996.

Abowd, G. D., Brotherton, J. A., and Bhalodai, J. "Classroom 2000: A system for capturing and accessing multimedia classroom experiences," in CHI'98, 1998.

Adwait, R. "A simple introduction to maximum entropy models for natural language processing," *IRCS Report* 97--08, University of Pennsylvania, 3401 Walnut Street, Suite 400A, Philadelphia, PA, May 1997.

Agius, H. W., and Angelides, M. C. "Developing knowledge-based intelligent multimedia tutoring systems using semantic content-based modeling," *Artificial Intelligence Review*, 13, pp55-83, 1999.

Alavi, M. and Leidner, D. E. "Review: Knowledge management and knowledge management systems: conceptual foundations and research issues," *MIS Quarterly* (25:1), March 2001.

Allan, J., Carbonell, J., Doddington, G., Yamron, J. and Yang, Y. "Topic detection and tracking pilot study: Final report," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

Allan, J., Lavrenko, V., Malin, D., Swan, R. "Detections, bounds, and timelines: UMass and TDT-3," In the *Proceedings of Topic Detection and Tracking Workshop* (TDT-3), 2000.

Ausubel, D.P. "The use of advance organizers in the learning and retention of meaningful verbal material," *Journal of Educational Psychology*, 51, 267-272, 1960.

Baltes, C. "The E-learning balancing act: training and education with multimedia," *IEEE Multimedia*, 8(4), 16-19, 2001.

Beeferman, D., Berger, A. and Lafferty, J. "Text segmentation using exponential models," in *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pp. 35-46, 1997 .

Blei, D. M. and Moreno, P. J. “Topic segmentation with an aspect Hidden Markov Model,” in *Proceedings of the 24th International Conference on Research and Development in Information Retrieval*, New York, NY: ACM Press, 2001.

Bouthemy, P., Dufournaud, Y., Fablet, R., Mohr, R., Peleg, S., Zomet, A. “Video hyper-link creation for content-based browsing and navigation,” *Workshop on Content-Based Multimedia Indexing*, CBMI'99, Toulouse, France, October 1999.

Callan, J.P. “Passage-level evidence in document retrieval,” In *Proc ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pp 302-309, Dublin, Ireland, 1994.

Cao, J., Crews, J.M., Lin, M., Burgoon, J.K., and Nunamaker, J.F. "Can People Be Trained to Better Detect Deception? Instructor-Led vs. Web-Based Training," in *Proceedings of the Ninth Americas Conference on Information Systems*, Tampa, Florida, 2003.

Cao, J. “Question answering on lecture videos: A multifaceted approach,” *Proceedings of the Fourth ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2004)*, Tucson, AZ, 2004.

Cao, J., Crews, J. M., Nunamaker, J. F. Jr., Burgoon, J. K., and Lin, M. “User experience with Agent99 Trainer: A usability study,” In *Proceedings of 37th Annual Hawaii International Conference on System Sciences (HICSS 2004)*, Big Island, Hawaii, 2004a.

Cao, J., Roussinov, D., Robles-Flores, J. A., and Nunamaker, J. F. Jr. “Automated question answering from videos: NLP vs. Pattern Matching,” *38th Annual Hawaii International Conference on System Sciences (HICSS 2005)*, Big Island, Hawaii, 2004b.

Cassell, J., Nakano, Y., Bickmore, T., Sidner, C., and Rich, C. "Annotating and generating posture from discourse structure in embodied conversational agents." *Workshop on Representating, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents*, *Autonomous Agents 2001 Conference*, Montreal, Quebec. May 29th, 2001.

Chaisorn, L., Chua, T., Koh, C., Zhao, Y., Xu, H., Feng, H. and Tian, Q. “A two-level multi-modal approach for story segmentation of large news video corpus,” Presented at *TRECVID Conference*, Published on-line at http://wwwnlpir. nist.gov/projects/tvpubs/tv.pubs.org.html, 2003.

Chen, B., Wang, H.-M. and Lee, L.-S. “Improved spoken document retrieval by exploring extra acoustic and linguistic cues,” *Proceedings of the 7th European Conference on Speech Communication and Technology*, 2001.

Choi, F. “Advances in domain independent linear text segmentation,” in *The North American Chapter of the Association for Computational Linguistics (NAACL)*, Seattle, USA, 2000.

Clarke, C., Cormack, G., Kisman, D., and Lynam, T. “Question answering by passage selection (Multitext experiments for TREC-9),” In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, 2000a.

Clarke, C., Cormack, G., and Tudhope, E. “Relevance ranking for one to three term queries,” *Information Processing and Management*, 36, pp 291–311, 2000b.

Cunningham, H. *Software Architecture for Language Engineering*. PhD Thesis, University of Sheffield, 2000.

Daft, R.L., and Lengel, R.H. “Organizational information requirements, media richness and structural design,” *Management Science* (32:5), pp. 554-571, 1986.

Das, M. and Liou, S. "A new hybrid approach to video organization for content-based indexing," *Int. Conf. on Multimedia Systems and Computing*, 1998 (Accidentally left out of proceedings).

Dorai, C., Kermani, P., and Stewart, A. “Elm-n: e-learning media navigator,” in *ACM Multimedia*, pp634–635, 2001.

Dorai, C., Oria, V., and Neelavalli, V. “Structuralizing educational videos based on presentation content,” in ICME’03, 2003.

Fujii, A., Itou, K., Akiba, T., Ishikawa, T. “A cross-media retrieval system for lecture videos,” *Proceedings of the 8th European Conference on Speech Communication and Technology* (Eurospeech 2003), Geneva, Switzerland, pp1149-1152, 2003.

Gargi, U., Kasturi, R., and Strayer, S. H. “Performance characterization of video-shot change detection methods,” *IEEE Transaction on Circuits Systems and Video Technology* 10(1), 1, 2000.

Glass, J., Hazen, T., Hetherington, L., and Wang, C. “Analysis and processing of lecture audio data: Preliminary investigations,” In *Proceedings of the HLT-NAACL 2004 Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval*, pp9-12, Boston, MA, May, 2004.

Halliday, M. and Hasan, R. *Cohesion in English*, Longman, 1976.

Harabagiu, S., et al. "FALCON: Boosting knowledge for answer engines," In *Proceedings of the Text Retrieval Conference* (TREC-9), 2000.

Hearst, M. A. "TextTiling: Segmenting text into multi-paragraph subtopic passages," *Computational Linguistics* (23:1), pp33-64, 1994.

Heinonen, O. "Optimal multi-paragraph text segmentation by dynamic programming," In *Proceedings of 17th International Conference on Computational Linguistics (COLING-A CL98)*, 1484-1486, 1998.

Hepple, M. "Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers, " In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, Hong Kong, 2000.

Hodge, G. "Systems of knowledge organization for digital libraries: Beyond traditional authority," Files: *Council on Library and Information Resources*, 2000, http://www.clir.org/pubs/reports/pub91/contents.html.

Hovy, E., Hermjakob, U., and Lin, C.-Y. "The use of external knowledge in factoid QA," In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 2001.

Hsu, W. H. M., and Chang S. F. "A statistical framework for fusing mid-level perceptual features in news video," Invited paper, ICME 2003, Baltimore, USA, 2003.

Ittycheriah, A., Franz, M., and Roukos, S. "IBM's statistical question answering system—TREC-10," In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 2001.

Johnson, S.E., Jourlin, P., Moore, G.L., Sparck Jones, K. and Woodland, P.C. "The Cambridge University spoken document retrieval system," *Proceedings of ICASSP*. 49-52. 1999.

Kan, M., Klavans, J.L. & McKeown, K. R. "Linear segmentation and segment significance," In *Proceedings of the 6th International Workshop of Very Large Corpora*, 197-205, 1998.

Kariya, S. "Online education expands and evolves," in *IEEE Spectrum*, May 2003.

Kaszkiel, M., Zobel, J. "Passage retrieval revisited," In *Proceedings of the Twentieth International ACM-SIGIR Conference on Research and Development in Information Retrieval*. Philadelphia. ACM Press, pp178-185, 1998.

Katz, S. M. "Distribution of content words and phrases in text and language modeling," *Natural Language Engineering*, 2(1), pp15-59, 1996.

Kaufmann, S. "Cohesion and collocation: Using context vectors in text segmentation," In *Proceedings of the 37th Annual Meeting of the Association of for computational Linguistics (Student Session)*, 591-595, College Park, USA, June. ACL, 1999.

Klavans, J. & Kan, M.Y. "Role of verbs in document analysis," In *COLING-ACL*, 680-686, 1998.

Knaus, D., Mittendorf, E., Schauble, P., and Sheridan, P. "Highlighting relevant passages for users of the interactive SPIDER retrieval system," In D.K. Harman, editor, *Proc. Text Retrieval Conference (TREC)*, pp 233- 243, Washington, 1995.

Kozima, H. & Furugori, T. "Similarity between words computed by spreading activation on an English dictionary," In *Proceedings of the European Association for Computational Linguistics*, 232—239, 1993.

Lee, G. G., Seo, J., Lee, S., Jung, H., Cho, B.-H., Lee, C., Kwak, B.-K., Cha, J., Kim, D., An, J., Kim, H., and Kim, K. "SiteQ: Engineering high performance QA system using lexico-semantic pattern matching and shallow NLP," In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, 2001.

Levy, E. T. *Communicating Thematic Structure in Narrative Discourse: The Use of Referring Terms and Gestures*. PhD thesis, University of Chicago, Chicago, Illinois, 1984.

Li, H. L., and Dong, A. J. "Hierarchical segmentation of presentation videos through visual and text analysis," submitted to *IEEE ISSPIT,* 2006.

Light, M., Mann, G. S., Riloff, E., and Breck, E. "Analyses for elucidating current question answering technology," *Journal of Natural Language Engineering, Special Issue on Question Answering*, Fall–Winter 2001.

Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., and Karger, D. R. "What makes a good answer? The role of context in question answering," In *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT-2003)*, 2003.

Lin, M., Crews, J. M., Cao, J., Nunamaker, J. F. Jr., & Burgoon, J. K. "AGENT99 Trainer: Designing a web-based multimedia training system for deception detection knowledge transfer," In *Proceedings of the Ninth Americas Conference on Information Systems*, Tampa, Florida, 2003.

Lin, M., Nunamaker, J. F. Jr., Chau, M. and Chen, H. "Segmentation of lecture videos based on text: a method combining multiple linguistic features," In *Proceedings of 37th Annual Hawaii International Conference on System Sciences (HICSS 2004)*, Big Island, Hawaii, 2004a.

Lin, M., Yuan, M. "Lecture video segmentation: A field study and case study," Project Report for MIS 578, 2004b.

Lin, M., Chau, M., Cao, J., and Nunamaker, J. F. Jr. "Automated video segmentation for lecture videos," *The International Journal of Technology and Human Interaction (IJTHI),* Athens, Greek: Idea Group Publishing, 2005a.

Lin, M., Diller, C., Meek, N. F., Huang Y., and J., Nunamaker, J. F. Jr. "Segmenting lecture videos by topic – From manual to automated method," *The Eleventh Americas Conference on Information Systems (AMCIS 2005)*, Omaha, Nebraska, 2005b.

LVCSR, AT&T Lab Speech Score. http://www.research.att.com/news/2002/June/LVCSR.html.

Liu, T., Hjelsvold R., and Kender, J. R. "Analysis and enhancement of videos of electronic slide presentations," in *Proceedings of International Conference on Multimedia and Expo (ICME)*, 2002.

Liu, T., and Kender, J. R. "Spatial-temporal semantic grouping of instructional video content," in *Proceedings of International Conference on Content-based Image and Video Retrieval (CIVR),* pp362-372, 2003a.

Liu, T., and Kender, J. R. "Lecture videos for e-learning: Current research and challenges," in *Proceedings of IEEE Sixth International Symposium on Multimedia Software Engineering,* pp.574-578*,* 2004.

Liu, Y. and Kender, J. R. "Fast video segment retrieval by sort-merge feature selection, boundary refinement and lazy evaluation," in *Computer Vision and Image Understanding*, pp147–175, 2003b.

McNeill, D., Quek, F., McCullough, K-E., Duncan, S., Furuyama, N., Bryll, R., Ma, X-F., and Ansari,.R., "Catchments, prosody, and discourse," *Gesture*, 2001.

Miller, G., Beckwith, R., Felbaum, C., Gross, D., and Miller, K. “Introduction to WordNet: An online lexical database,” *International Journal of Lexicography* (3:4, pp235-312), 1990.

Mittendorf, E., and Schauble, P. “Document and passage retrieval based on hidden Markov models,” In *Proc. ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pp 318-327, Dublin, Ireland, 1994.

Moldovan, D., et al., “LCC tools for question answering”, in L.P. Buckland and E. Voorhees (eds): *Proc of TREC 2003*, NIST, Gaithersburg, USA, 2003.

Morris, J., and Hirst, G. “Lexical cohesion computed by thesaural relations as an indicator of the structure of text,” *Computational Linguistics*, 17, 21-48, 1991.

Mukhopadhyay, S., and Smith, B. “Passive capture and structuring of lectures,” in *ACM Multimedia*, pp477–487, 1999.

Ngo, C. W., Pong T. C., and Chin, R. T. “ Video partitioning through temporal slices analysis, *IEEE Trans. on Circuits and Systems for Video Technology* , Vol. 11, No. 8, pp941-953, 2001.

Ngo, C. "Recent advances in content-based video analysis", *International journal of image and graphics*, Vol. 1, Issue 3, p445, ISSN: 0219-4678, 2001.

Ngo, C. W., Wang F., Pong, T. C. “Structuring lecture videos for distance learning applications,” in *Proceedings of Fifth International Symposium on Multimedia Software Engineering*, pp. 215-222, 2003.

Dulce, B., Ponceleon, and Srinivasan, S. “Structure and content-based segmentation of speech transcripts,” *SIGIR 2001*, pp 404-405, 2001.

Ponte, J. M., and Croft, W. B. “Text segmentation by topic,” In *European Conference on Digital Libraries*, pp 113-125, Pisa, Italy, 1997.

Porter, M. “An Algorithm for Suffix Stripping,” *Program* (14:3, pp. 130-137), 1980.

Quek, F., McNeill, D., Bryll, R., Kirbas, C., Arslan, McCullough, K-E., Furuyama, N., and H., Ansari, R. “Gesture, Speech, and Gaze Cues for Discourse Segmentation,” in *proceeding of IEEE Conference on Computer Vision and Pattern Recognition* (2), pp. 247–254, 2000.

Ramshaw, L., and Marcus, M. “Text chunking using transformation-based learning,” In *Proceedings of the Third ACL Workshop on Very Large Corpora*, MIT, June, 1995.

Reynar, J. C. “An automatic method of finding topic boundaries”, in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (Student Session)*, Las Cruces, New Mexico, pp. 331-333, 1994.

Reynar, J. C. *Topic Segmentation: Algorithms and Applications*, PhD thesis, Computer and Information Science, University of Pennsylvania, 1998.

Rowe, L. A., and Gonzlez, J. M. “BMRC lecture browser demo,” In *WWW document, URL* http://bmrc.berkeley.edu/frame/projects/lb/index.html, 2000.

Rui, Y., Gupta, J. A., Grudin, and He, L. “automating lecture capture and broadcast: technology and videography,” *Multimedia Systems*, 2004.

Salton, G., Allan, J., and Buckley, C. “Approaches to passage retrieval in full text information systems,” In *Proc. ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pp 49-58, Pittsburg, 1993.

Salton, G., Singhal, A., Buckley, C. and Mitra, M. “Automatic text decomposition using text segments and text themes,” in *Proceedings of Hypertext'96*, ACM Press, New York, pp. 53-65, 1996.

Scott, S., and Gaizauskas, R. “University of Sheffield TREC-9 Q & A System”, in D. Harman and E. Voorhees (eds): *Proc. of TREC 2000*, NIST, Gaithersburg, USA, 2000.

Shriberg, E., Stolcke, A., Hakkani-Tur, D. and Tur, G. “Prosody-Based automatic segmentation of speech into sentences and topics,” *Speech Communication (Special Issue on Accessing Information in Spoken Audio)* 32:1-2, pp. 127-154, 2000.

Slaney, M. and Ponceleon, D. “Hierarchical segmentation: finding changes in a text signal,” In *Proc. of the SlAM Text Mining 2001 Workshop*, Chicago, IL, pp. 6-13, 2001.

Solomon, P. “Exploring structuration in knowledge organization: Implications for managing the tension between stability and dynamism,” *Advances in Knowledge Organization*, 7, pp. 254-260, 2000.

Soubbotin, M. M., and Soubbotin, S.M. “Use of patterns for detection of answer strings: A systematic approach”, in L.P. Buckland and E. Voorhees (eds): *Proc. of TREC 2002*, NIST, Gaithersburg, USA, 2002.

Stanford Online, http://scpd.stanford.edu/scpd/students/onlineclass.htm.

Stokes, N. "Applications of lexical cohesion analysis in the topic detection and tracking domain," Department of Computer Science, University College Dublin, April 2004.

Smith, T., Ruocco, A., and Jansen, B. "Digital video education," in *ACM SIGCSE'99*, pp. 122–126, 1999.

Syeda-Mahmood, T., and Srinivasan, S. "Detecting topical events in digital video," in ACM Conference on Multimedia, 2000.

Tellex, S., Katz, B., Lin, J., Marton, G., and Fernandes, A. "Quantitative evaluation of passage retrieval algorithms for question answering," In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, July, 2003.

Thong, V., Moreno, J.-M., Logan, P.J., Fidler, B., Maffey, B., and Moores, M. "Speechbot: an experimental speech-based search engine for multimedia content on the web," *IEEE Trans on Multimedia*, 4(1), pp 88-96. 2002.

Tolle, K. and Chen, H. "Comparing noun phrasing techniques for use with medical digital library tools," *Journal of the American Society for Information Science (Special Issue on Digital Libraries)* (51:4, pp. 352-370), 2000.

Tur, G., Hakkani-Tur, D., Stolcke, A., Shriberg, E. "Integrating prosodic and lexical cues for automatic topic segmentation," *Computational Linguistics* (27:1, pp. 31-57), 2001.

Utiyama, M., Isahara, H. "*A statistical model for domain - independent text segmentation*," In the *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, (EACL-01), pp. 491-498, 2001.

Voorhees, E.M. "Overview of the TREC 2002 question answering track," In *notebook of the Eleventh Text Retrieval Conference (TREC'2002)*, pp 115-123. 2002.

Voutilainen, A. "Helsinki taggers and parsers for English," In J. M. Kirk (Ed.) *Corpora Calore: Analysis and Techniques in Describing English*. Rodopi, Amsterdam & Atlanta, 2000.

Wactlar, H. D. "Informedia - search and summarization in the video medium," in *Proceedings of Imagina 2000 Conference*, Monaco, 2000.

Wang, F., Ngo, C. W., and Pong, T. C. "Exploiting self-adaptive posture-based focus estimation for lecture video editing," *ACM Multimedia Conference*, Nov 2005.

Wang, F., Ngo, C. W., and Pong, T. C. "Gesture tracking and recognition for lecture video editing," *Int. Conf. on Pattern Recognition*, 2004.

Wang, G. Chua, T.S., Wang, Y.C. "Extracting key semantic terms form Chinese speech queries for Web searches," 41st Annual Meeting of the Association of Computational Linguistics (ACL'03), Sapporo, Japan, pp 248-255, July 2003.

Wllkinson, R. "Effective retrieval of structured documents," In *Proc. ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pp 311-317, Dublin, Ireland, 1994.

Wu, L., et al. "FDUQA on TREC2003 QA task," in L.P. Buckland and E. Voorhees (eds): *Proc. of TREC 2003*, NIST, Gaithersburg, USA, 2003.

Yaari, Y. "Segmentation of expository texts by hierarchical agglomerative clustering," In *Proceedings of Recent Advances in Natural Language Processing*, Bulgaria, 1997.

Yaari, Y. *Intelligent Exploration of Expository Texts*, PhD thesis, Bar-Ilan University, Ramat-Gan, Israel, 1999.

Yamamoto, Y. A. N., and Ogata, J. "Topic segmentation and retrieval system for lectures videos based on spontaneous speech recognition," *EUROSPEECH Geneva*, 2003.

Yamron, J. P., Carp, I., Gillick, L., Lowe, s., and van Mulbregt, P. "Topic tracking in a news stream," in *Proceedings of the DARPA Broadcast News Workshop*, pp. 133-136, 1999.

Yang, H., Chaison, L., Zhao, Y., Neo, S.-Y., and Chua, T.-S. "VideoQA: Question answering on news video," In the *Proceedings of the Eleventh Annual ACM International Conference on Multimedia (ACMM'2003)*, Berkeley, California, USA, 2-8 Nov 2003.

Youmans, G. "A new tool for discourse analysis: The vocabulary management profile," *Language* (67:4), 1991, pp. 763-789.

Zhang, D. S. *Virtual Mentor and Media Structuralization Theory*, PhD thesis, University of Arizona, Tucson, AZ, 2002.

Zhang, D., and Nunamaker, J. "A natural language approach to content-based video indexing and retrieval for interactive e-learning," *IEEE Transactions on Multimedia*, Volume 6, Number 3, pp. 450-458, 2004.

Zhang, H. J. and Smoliar, S. W. “Developing power tools for video indexing and retrieval,” in *Proceedings of SPIE'94 Storage and Retrieval for Video Databases*, San Jose, CA, USA, 1994.

Zhang, W. “Multimedia, technology, education and learning,” *Technological Innovations in Literacy and Social Studies Education*, The University of Missouri-Columbia, 1995.

Zobel, J., Moffat, A., Wilkinson, R., and Sacks-Davis, R. “Efficient retrieval of partial documents,” *Information Processing & Management*, 31(3), pp 361-377, 1995.