LEARNING ON DEMAND - COURSE LECTURE DISTILLATION BY INFORMATION EXTRACTION AND SEMANTIC STRUCTURING FOR SPOKEN DOCUMENTS

Sheng-yi Kong, Miao-ru Wu, Che-kuang Lin, Yi-sheng Fu, Lin-shan Lee

Speech Lab, College of EECS
National Taiwan University, Taipei, Taiwan, Republic of China
sykong@speech.ee.ntu.edu.tw,lslee@gate.sinica.edu.tw

ABSTRACT

This paper presents a new approach of organizing the course lectures (as spoken documents) for efficient learning on demand by the users. By properly matching the course lectures with the slides used, we divide the course lectures into hierarchical "major segments" with variable length based on the topics discussed. Key term extraction, hierarchical summarization and semantic structuring are then performed over these "major segments". A key term graph is also constructed, based on which the various major segments of the course can be linked. In this way, the user can ask questions to the system, and develop his own road map of learning the knowledge he needs considering his available time and his background knowledge, based on the semantic structure provided by the system. A preliminary prototype system has been successfully developed with encouraging initial test results.

Index Terms— Course lectures, Spoken documents, Semantic structuring, Key term hierarchy

1. INTRODUCTION

With fast advances of spoken document processing technologies, it is now possible to manage huge quantities of multimedia information based on the included audio information[1], because the speech information associated with the multimedia is usually the key for extracting the information. Broadcast programs, meeting records and course lectures are typical examples for possible application areas. Course lectures have been a focus of research along this direction, not only because life-long learning has become necessary in the era of knowledge explosion, but the ever-increasing bandwidth of Internet and continuously falling costs for memory have made it possible to distribute huge quantities of complete course lectures worldwide very easily [2, 3, 4].

A major difficulty of efficiently utilizing the many complete course lectures available over the network is that it takes quite long time to listen to a complete course (e.g. a complete course may include 45 hrs.), and it may not be easy for leaders or researchers working in the industry to spend so much time to learn a complete course. On the other hand, the content of a course is usually well structured; the learner cannot understand an advanced subject without knowing related fundamentals of the background. As a result, direct retrieval of the course content for some advanced subjects is usually not helpful to the learner, simply because the retrieved results are difficult to understand. Also, after learning a subject the learner usually doesn't know what the related subjects are which should be learned next.

In this paper, we present a new approach of managing the course lectures. By careful information extraction and semantic structuring for the spoken documents (speech information) associated with the course lectures, we are able to construct the semantic structure for the lectures of a course. This includes dividing the course lectures into "major segments" based on topics discussed; key term extraction, hierarchical summarization, semantic structuring for the major segments; and a key term graph to link all the major segments. The user can thus ask questions to the system and learn what he needs in his own way. This is referred to as "Learning on Demand - Course Lecture Distillation" in this paper. A preliminary prototype system has been successfully developed at National Taiwan University (NTU), which is referred to as NTU Virtual Instructor in this paper.

2. LEARNING ON DEMAND - COURSE LECTURE DISTILLATION

A few examples are shown in Fig. 1 for user/system interactions with the concept presented here. They explain what we mean by "Learning on Demand" or "Course Lecture Distillation". The example course used here is on Digital Speech Processing. In Fig. 1(a), the user asks about *language modeling*. The system replies that this phrase appears in segment 1.3.5 (Chapter 1, slide No.3, the 5-th section) with subtitle "Language Model", segment 2.7 (chapter 2, slide No.7), and segment 6.0 (chapter 6) with title "Language Model", etc.. These are three major segments with length ranging from 2.1 mins up to 235.9 mins. In each case the user may click to view the slides, to listen to the summary, or to see if he understand all the key terms involved. All these are to help the user to decide whether he should click to listen to the segments or not. He can click to listen to the segment, or to start with the slide or chapter beginning if he wishes. In this example the question is relatively fundamental, and the user can learn from the beginning and then enter more details and more advanced parts later on. This is referred to as "forward learning".

In Fig. 1(b), the user asks about *triphone*. When he clicks the key terms involved, he realizes he doesn't understand a key term called *decision tree*. He may select to click the segments offered by the system regarding *decision tree* first, to make sure he can understand the segment on *triphone*. In the course lectures *decision tree* was actually mentioned before *triphone*. This is referred to as "backward learning". In either forward or backward learning, the system also offers a list of "other related subjects" or "other related key terms" to help the user during learning.

3. PROPOSED APPROACH

The proposed approach to realize the above concept is summarized in this section.

(a)
- User: Tell me about Language Modelling.
- System: *Language Modeling* was mentioned in

1.3.5 Language Model (2.1 mins)
View Slides | Summary | Key Terms | Start with the slide beginning | Start with the chapter beginning

2.7 Language Model - N-gram (23.9 mins)
View Slides | Summary | Key Terms | Start with the chapter beginning

6.0 Language Modeling (235.9 mins)
View Slides | Summary | Key Terms

Other Related Subjects

(b)
- User: What is a Triphone?
- System: *Triphone* was mentioned in

5.10 Training Triphone Models (16.9min)
View Slides | Summary | Key Terms | Start with the chapter beginning

- User: Click Key Terms
- System: Key terms for *5.10 Training Triphone Models* are

Decision Tree was mentioned in
5.7 Decision Tree (38.6 mins)
View Slides | Summary | Key Terms | Start with the chapter beginning

Node Splitting was mentioned in
⋮

Other Related Key Terms

Fig. 1. Examples of user/system interaction for course lecture distillation with (a) forward learning and (b) backward learning (5.7 was offered before 5.10)

3.1. Overview of The Proposed Approach

An overview of the proposed approach is shown in Fig. 2. On the upper left corner the course lectures includes 3 parts: audio, video and slides. Audio and video signals have synchronized time indices, but the slides are not synchronized with the signals in general (although there exist software tools to synchronize the slides as well, we didn't assume such tools were used). Audio segmentation was first performed using acoustic information to divide the audio signal (and therefore video signal) into "short segments" (roughly 1-6 utterances long) for lecture transcription, or the large vocabulary continuous speech recognition on the right of the figure. Special efforts have to be made here to handle the spontaneous nature of the lecture speech.

The core of the proposed approach is on the lower left part of Fig. 2, content organization and retrieval. Topic segmentation is to try to collect a number of "short segments" which discuss the same subject topic together into "major segments", primarily based on the transcribed short segments and the slides. The "major segments" have variable lengths. It may cover a section of a slide, a whole slide, or even a chapter of many slides. These major segments form a segment hierarchy, with an example partial list shown in Fig. 3. Semantic analysis is then performed on the major segments on the lowest level, or each leaf node of the segment hierarchy. Key terms are then extracted. Hierarchical summarization is performed on all major segments, longer or shorter on higher or lower levels of the segment hierarchy. So every node on the segment hierarchy has a summary. Semantic structuring is primarily based on a key term graph constructed for all the key terms extracted with an example shown in Fig. 4, which is used to link all the major segments semantically. The basic unit for information retrieval is the short segments,

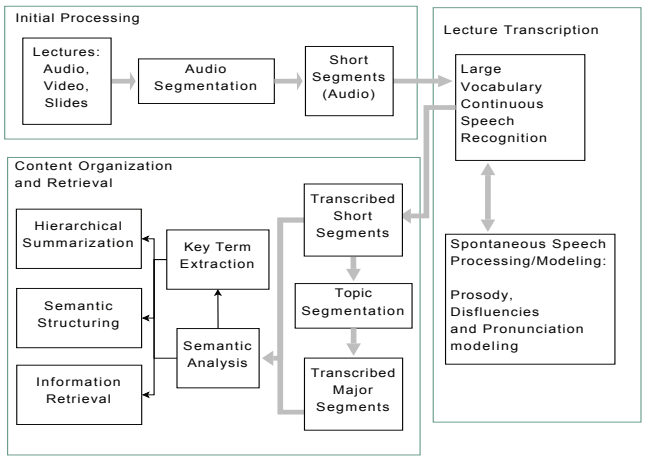

Fig. 2. Overview of the proposed approach.

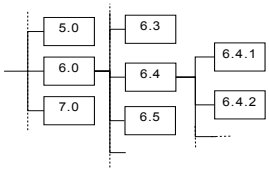

Fig. 3. A partial list of the segment hierarchy. 6.0 is chapter 6, 6.4 is the 4-th slide of chapter 6 , and 6.4.1 is the first section of that slide.

although the major segments can also be retrieved.

3.2. Topic Segmentation

Topic Segmentation here is performed by a decoding process over Hidden Markov Models (HMMs) of the slides and their sections. If a slide has several sections each under a subtitle, each section is modeled as an HMM state. If not, the whole slide is modeled as an HMM state. Given a sequence of slides (or their sections) modeled as a sequence of states, $\{s_j, j = 1, 2, \ldots, M\}$, and a sequence of short segments taken as a sequence of observations, $\{o_i, i = 1, 2, \ldots, T\}$, the probability for observing o_i in state s_j is evaluated by

$$P(o_i|s_j) = \prod_{l=1}^{3} \left[\frac{\sum_{w_k \in o_i, w_k \in s_j} (1-\epsilon_k) c_j(w_k)}{\sum_{w_k \in s_j} (1-\epsilon_k) c_j(w_k)} \right]^{\alpha_l}, \quad (1)$$

$$\epsilon_k = -\frac{1}{\log T} \sum_{i=1}^{T} [\frac{c_i(w_k)}{m_k}] \log [\frac{c_i(w_k)}{m_k}], \quad (2)$$

where w_k is a word pattern ($l = 1$ for a single word such as "speech", $l = 2$ for a 2-word pattern such as "Viterbi algorithm", $l = 3$ for a 3-word pattern), $c_i(w_k)$ and $c_j(w_k)$ are respectively the counts of the word pattern w_k in the observation o_i and and the state s_j, α_l are weight parameters, m_k is the total counts of w_k in all

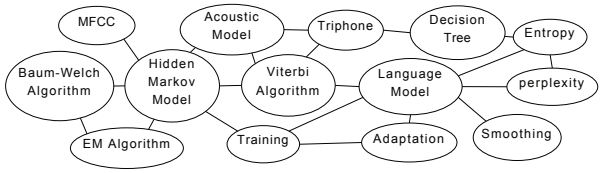

Fig. 4. An example of Key Term Graph.

the observations, T is the total number of observations. So ϵ_k is the normalized entropy for word pattern w_k indicating the importance of the word pattern, and $P(o_i|s_j)$ is simply the total counts of all word patterns existing in both o_i and s_j, properly weighted , divided by the total counts of all word patterns in s_j, averaged over 3 different sizes of word patterns. By assigning state transition probabilities and initial probabilities, the topic segmentation is achieved by Viterbi decoding.

3.3. Semantic Analysis

The semantic analysis is primarily based on Probabilistic Latent Semantic Analysis (PLSA)[5]. Here each leaf node of the segment hierarchy is taken as a spoken document d_i. Assume there are a total of N such documents (major segments) for the course, $\{d_i, i = 1, 2, \ldots, N\}$, which includes a total of L terms, $\{t_j, j = 1, 2, \ldots, L\}$. A set of latent topic variables is defined, $\{T_k, k = 1, 2, \ldots, K\}$, to characterize the "term-document" co-occurrence relationships. The conditional probability of a document d_i generating a term t_j thus can be parameterized by

$$P(t_j|d_i) = \sum_{k=1}^{K} P(t_j|T_k)P(T_k|d_i). \quad (3)$$

The PLSA model can be optimized with EM algorithm by maximizing a likelihood function [5].

3.4. Key Term Extraction

Key terms are extracted based on the slide information and the spoken document transcriptions jointly. From the slides, after stop word removal, some rule-based as well as statistical approaches were used to extract key terms(words and phrases) using such features as term frequencies, inverse slide frequencies (idf but taking slides as documents), the word pattern entropy in eq(2) above, sentence length, words in title/subtitle, letters in upper case, etc.. For the spoken documents, it has been found that the latent topic entropy evaluated based on PLSA parameters have been very helpful in extracting key terms [6]. The latent topic entropy of a term t_j is evaluated from the latent topic distribution $\{P(T_k|t_j), k = 1, 2, \ldots, K\}$ of the term obtained from PLSA and defined as:

$$\mathrm{EN}(t_j) = -\sum_{k=1}^{K} P(T_k|t_j) \log P(T_k|t_j). \quad (4)$$

Clearly, a lower entropy implies the term carries more topical information for a few specific latent topics, thus is more significant semantically. Thesse two different directions can then be integrated.

Very often a phrase of a few words is a more specific key term (e.g. "language model" or "information theory" are more specific than "language" or "model" or "information"). Therefore the co-occurrence statistics of the key terms extracted above are further calculated and stored in an efficient data stucture called "PAT-Tree" [7]. With the forward and backward co-occurrence PAT-Trees, some of the above extracted key terms such as "language" and "model" are further combined into a new key term (or actually a key phrase).

3.5. Hierarchical Summarization

A summary in speech form (together with the associated video and slides) is generated for each node (major segment) on the segment hierarchy, longer or shorter on higher or lower levels. The speech form summary is generated in general in the same way as was done in previous work [8], i.e. utterances are selected based on some scores, including the latent topic entropy as in eq. (4) for the terms in the utterances, and then concatenated to form the summary. A major difference here is that those utterances with transcriptions closely matched to the titles of the slides and the subtitles of the sections in the slides are given higher scores. Also, the length of the summary never exceeds an upper limit regardless of how long the major segment is. For example, if a major segment correspond to a chapter of 3 hrs long, the summary is still within 50 sec so the user can have a quick understanding regarding what the chapter is about, in order to decide how he should listen to the lectures. In such case the summarization ratio is very low. Very often in such cases the summary includes primarily those utterances closely matched to the title and subtitles of the slides.

3.6. Semantic Structuring

This is the core of the proposed approach, to construct a key term graph to link semantically all major segments (higher or lower in the segment hierarchy) and the concepts discussed in the course lectures, such that the system can guide the user to listen to the related major segments one by one, either forward or backward.

Relationships between key terms (including key phrases) are evaluated based on the co-occurrence statistics as well as the probabilities obtained from PLSA. The key term graph is then constructed based on such relationships after deletion of noisy terms and weak relationships.

3.7. Information Retrieval

The information retrieval approach used here is kind of standard. The only difference is that here the retrieved documents can be any short segment or any node on the segment hierarchy. The retrieved documents can not only be the major segments on the leaf nodes of the segment hierarchy, but can also be a whole chapter if the query term is included in the title of the chapter, or a whole slide if the query term is included in the title of the slide, etc..

4. THE COURSE LECTURE CORPUS USED AND THE PRELIMINARY PROTOTYPE SYSTEM

A single course lecture corpus was used for the preliminary prototype system, which is a course on Digital Speech Processing with a total length of 45 hrs. It was offered in National Taiwan University (NTU) by a single instructor in Mandarin Chinese, while all the terminologies were produced directly in English, not translated into Chinese, but inserted in the Chinese utterances just as Chinese words. Such a code-switching environment is quite usual in courses offered in Taiwan. The slides used are completely in English.

An integrated Mandarin/English phone set of 75 units including shared units was used to train triphone acoustic models. A combined Chinese/English lexicon of 12.3k words was used, in which all words used in the English slides (special terminologies, key terms and general words) were included. Class-based tri-gram language models were trained using both topic-related and style-related corpus (2.79 million words), then properly adapted by the slide information. The initial recognition accuracies ranged from 64.77% to 81.63% for different major segments for Chinese characters, and was 66.77% for English words. These accuracies were not high since various techniques to improve the acoustic/linguistic models and to handle the spontaneous speech and disfluencies were not performed yet.

A preliminary prototype system has been successfully developed. It is referred to as NTU Virtual Instructor here. The block diagram of the system is as shown in Fig. 2 using proposed approaches mentioned in section 3.

5. SOME INITIAL EXPERIMENTAL RESULTS

5.1. Initial Results on Topic Segmentation

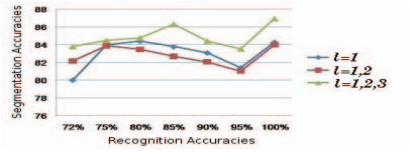

Fig. 5. Initial results on topic segmentation.

The performance of topic segmentation was evaluated with respect to reference segmentation by students taking the course. The computer-generated segment boundaries within 2 short segments from the reference boundaries were taken as correct segmentation. By randomly selecting some recognition errors replaced by reference transcriptions, we were able to obtain results for different recognition accuracies, if further improvements in recognition can be achieved in the future. The results of segmentation accuracy were shown in Fig. 5, from which we see using word patterns w_k of size $l = 1, 2, 3$ together gave the best results of roughly 84%-87%, and this number is not too much dependent on the recognition accuracy as long as the recognition accuracy is reasonably high.

5.2. Initial Results on Key Term Extraction

The reference key terms were extracted by 2 students in National Taiwan University who took the course. The performance of the key term extraction approaches was evaluated with respect to the reference key terms for chapters 4 to 17 in terms of precision/recall and F-measure. The results are shown in Fig. 6. High recall (between 77% and 96%) but relatively low precision was obtained , which gave moderate F-measures. From Fig. 6 it seems the performance for the present approach is topic dependent, i.e., mathematics or fundamental oriented chapters (chpaters 4-7) had very low precision, but concept or application oriented chapters (others) had higher precision. Clearly more investigations are needed here.

5.3. Initial Results on Hierarchical Summarization

The reference summaries were produced by two people, the course instructor and a teaching assistant. They simply selected a set of utterances meeting a desired summarization ratio to be taken as the reference summaries. Two levels of summaries were extracted, one for a single slide, the other for a whole chapter. Only a summarization ratio of 10% was evaluated. The work for other cases are still

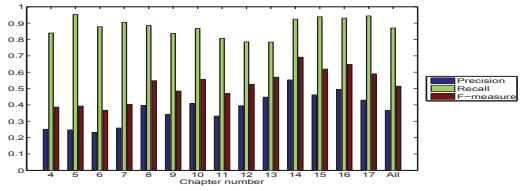

Fig. 6. Initial results on key term extraction.

Table 1. The ROUGE-1/ROUGE-L F-measures (%) of hierarchical summarization for (1) using spoken information alone and (2) plus slides information.

	Slides			Chapters		
approaches	SS	LTS	LTE	SS	LTS	LTE
(1)Spoken	52.2/49.0	52.1/49.5	54.2/49.1	63.0/61.7	64.2/62.8	68.6/66.9
(2)Plus slides	53.2/50.3	54.2/51.4	54.5/51.4	64.6/63.4	65.2/64.0	68.7/67.0

in progress. The previously proposed summarization methods based on significance score(SS) [9], latent topic significance (LTS) and latent topic entropy (LTE) [8] were evaluated by the well-known summarization evaluation package ROUGE [10]. The results are listed in Table 1. Two cases were evaluated: (1) using the spoken documents alone and (2) exploiting the information in the corresponding chapter slides, as in rows (1) and (2) in Table 1 with both ROUGE-1 and ROUGE-L scores listed. The results in Table 1 indicate that reasonable summarization performance is achievable using all the three approaches SS, LTS, and LTE with minor differences, and LTS seems to be slightly better. The slide information offered some but limited help, although not in all cases. Clearly more investigations are needed here.

6. CONCLUSION

This paper presents a new approach of organizing the course lectures for efficient "learning on demand" by the users. By careful information extraction and semantic structuring, we tried to construct the semantic structures of a course. A preliminary prototype system has been developed with some encouraging initial test results. More investigations are definitely still needed though.

7. REFERENCES

[1] L.-s. Lee, S.-y. Kong, Y.-c. Pan, Y.-s. Fu, and Y.-t. Huang, "A multi-layered summarization of spoken document archives by information extraction and semantic structuring," in *In Proceedings of Interspeech*, 2006.

[2] W. Hrst, T. Kreuzer, and M. Wiesenhtter, "A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web," in *In Proceedings of IADIS WWW/Internet 2002 Conference*, 2002, pp. 135–143.

[3] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the mit spoken lecture processing project," in *In Proceedings of Interspeech*, 2007.

[4] S. Togashi and S. Nakagawa, "A browsing system for classroom lecture speech," in *In Proceedings of Interspeech*, 2008.

[5] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of the 15th Conference on Uncertainty in AI*, 1999.

[6] Y.-c. Hsieh, Y.-t. Huang, and C.-c. Wang, "Improved spoken document retrieval with dynamic key term lexicon and probabilistic latent semantic analysis (plsa)," in *In Proceedings of ICASSP*, 2006.

[7] Lee-Feng Chien, "Pat-tree-based keyword extraction for chinese information retrieval," in *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1997, pp. 50–58, ACM.

[8] S.-y. Kong and L.-s. Lee, "Improved summarization of chinese spoken documents by probabilistic latent semantic analysis (plsa) with further analysis and integrated scoring," in *Proc. of IEEE/ACL Workshop on Spoken Language Technology*, 2006.

[9] M. Hirohata, Y. Shinnaka, K. Iwano, and S. Furui, "Sentence extraction-based presentation summarization techniques and evaluation metrics," in *Proc. of ICASSP*, 2005, pp. SP–P16.14.

[10] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. of Workshop on Text Summarization Branches Out*, 2004, pp. 74–81.