published as: Stephan Repp, Christoph Meinel: Semantic Indexing for Recorded Educational Lecture Videos;
Proceedings of the 4th Annual IEEE International Conference on Pervasive Computing and Communications Workshops (PerComW 2006);
Pisa (Italy), pp.240-245.

Semantic Indexing for Recorded Educational Lecture Videos

Stephan Repp, Christoph Meinel
Hasso-Plattner-Institut für Softwaresystemtechnik GmbH
University of Potsdam
D-14440 Potsdam, Germany
{repp,meinel}@hpi.uni-potsdam.de

Abstract

In this paper, we present a general architecture and a new retrieval method for an educational system that is based on a knowledge base of existing recorded lectures. The extraction of metadata from the multimedia resources is one of the main parts of this paper. The recorded lectures are transcripted by an out-of-the-box speech recognition software. The speech recognition software generates a time stamp for each word. These resources are divided into clusters, marked by timestamps, so that search engines are able to find the exact position of particular information inside a course. Finally, a retrieval method is presented that allows users to find "example", "explanation", "overview", "repetition", "exercise" for a particular word or topic-word. This is useful for the student's learning process and allows easy navigation of the multimedia database for pervasive learning.

1. Introduction

In the past decade, we have witnessed a dramatic increase in the availability of online academic lecture material. For instance, the newly developed tele-teaching system called "tele-Task" [1][2]. This system allows a video sequence (for example the teacher at the blackboard) to be bundled with the capture of the speakers desktop. This multimedia stream can be broadcast live or on demand over the Internet. These archived educational resources can potentially change the way people learn. For example students with disabilities can improve their educational skills, and professionals can follow recent advances in their field. In fact, such recorded lecture videos are suitable for distance learning, and can be seen as a complement to normal classroom courses [11]. However, there are two technical problems in the use of recorded lectures for pervasive learning: the problem of easy access to the content of multimedia lecture videos and the problem of finding the semantically appropriate information very quickly.

An e-learning or tele-teaching tool may not be useful for pervasive learning, if the information search makes large demands on bandwidth, memory and cpu-time because of the limited mobile hardware.

In this paper, we present our efforts in putting together results from different fields and projects in order to create a general architecture for practical learning that promote independent e-learning. It can be used for distant learning, adult education or in classical school courses. Our solution has the following properties:

- The benefits gained the large potential of knowledge in archives of recorded lectures
- Retrieval of information and additional meta information for the student (we called them *meta-phrase)* such as "example", "overview", "definition" and "repetition" for the student
- Easily integrated algorithm and techniques for e-learning systems

2. Related Work

Little information can be found in literature about educational systems that use a semantic search engine for finding additional information (meta-phrase) effectively in a knowledge base of recorded lectures. In this section, we briefly present three related projects concerning lecture video segmentation and indexing.

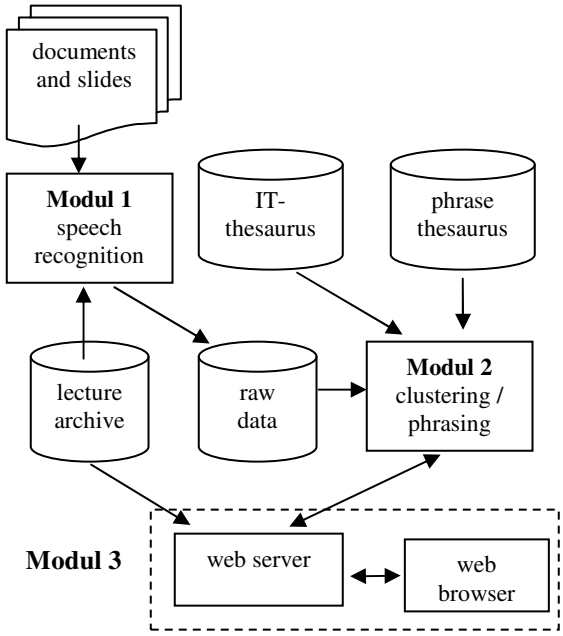
The design and adaptation of an automatic video browsing and retrieval system is presented in [3]. It describes a speech recognition module that recognizes speech in scenes, and an extraction module that extracts the texts from key frames, and then constructs the textual indices for retrieval. Unfortunately, the system is not adapted to lecture videos, no semantically

search is possible and the influence of the word error rate (WER) is not covered in this work. In [4] a high vocabulary automatic speech recognition system (commercial system, "out-of-the-box") to index recorded lectures is evaluated. However the accuracy of the speech recognition software is rather low – the recognition accuracy of audio lecture is approximately 22%-60%. It is shown in [5] that audio retrieval can be performed with out-of-the-box speech- recognition-software.

3. Overview of our architecture

In this section, we present a schematic overview of our architecture for educational systems. It is highly modularized in three basic functional modules (see figure 1).

Figure 1. Architecture of the system



MODULE 1: Lectures are recorded in a multimedia form, for example as *RealMedia* files (`.rm`) like in our implementation. The conversion of the audio-data into text, and the pre-processing of that text is the task of module 1. A speech recognition system is used to generate text data with a time-stamp on each word. Documents and slides are used for adding new words to the speech-recognition-software. The transcript data with the time stamps is stored in a relational database and we call this data "*raw data*".

MODULE 2: The automatic clustering, phrasing and extraction of required video sequence are the tasks of module 2. In fact, the raw data from module 1 and that from both of thesaurus act as the input data for this module. We use an IT specific thesaurus and a thesaurus for detecting "example", "overview", "repetition", and "definition/explanation". These patterns are called "*meta-phrase"*. This module is the main part of this paper and is presented in chapter 5.

MODULE 3: The well-known architecture consists of a web-server and an ordinary web-browser. We generate smile-files with PHP-scripts to create a user-friendly environment. This means that we use dynamically generated Smile-files for adapting the desktop settings (for example the resolution of the display) and for adapting the web-browser used by the students. There is a wide variety of hardware that may be adapted to this, for instance smart-phones, hand-helds and other mobile-computers with internet access.

4. Speech recognition and WER evaluation

An interesting quality study towards using large vocabulary automated speech recognition systems to index recorded presentations was published in [4]. The authors analyzed open-source speech-recognition-software, which provide an out-of-the-box speech recognition engine for German. Unfortunately, they could not identify open source software that fulfilled all their requirements for an easy to use speech recognition system. Although many of those systems can be configured and optimized for a given situation to expand the accuracy of their engine, these tuning options are nevertheless in most cases reserved for developers. Hence, after evaluating several state-of-the-art speech recognition systems, it was decided that only one commercial large vocabulary automatic speech recognition engine be considered for further experiments. We used the *Word Error Rate (WER)* and the *Word Accuracy (WA)* as a measure for speech recognition performance [6].

The recorded tele-TASK lectures "*Technische Grundlagen des WWW*" from the first semester 2005, which can be found at *www.tele-task.de* are used. We integrated the domain specific words to the speech recognition software from the existing power point slides of the lecture. The lecturer uses a common analogous line-in microphone to train the speed recognition software for 15 minutes. We compute a WA between 60% and 80%, depending on the acoustic conditions in the lecture room (background noise and overall acoustics). Our results for the WER are identical to those found in the literature [7].

It seems evident that performing a search on converted documents that are not error free will not

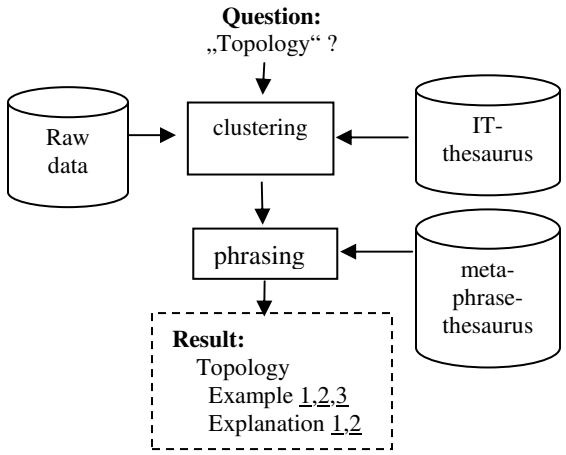
produce satisfactory results. Multimedia documents of live-recorded lectures consist of unscripted and spontaneous speech. Thus, lecture data has much in common with casual, or natural speech data, including false starts, extraneous filler words (such as "okay" and "well"), and non-lexical filled pauses (such as "uh" or "um") [9]. One can also easily observe that the colloquial nature of the data is dramatically different in style from the same presentation of this material in a textbook. In other words, the textual format is typically more concise and better organized.

In [10] they calculate a redundancy of about r~0.73 for the German language. In other words, in the German language 27% of the language carries information, the other 73% requires error detection and correction. This means that it is theoretically possible to get all required information from our raw data.

5. Clustering and semantic detection

The process of the metadata extraction is following: The query word of the user expanded with the word of the IT-Thesaurus. With this subset of patterns and words the so called *cluster algorithm* is started. In Chapter 5.1 the cluster algorithm is explained.

Figure 2. Clustering and phrasing



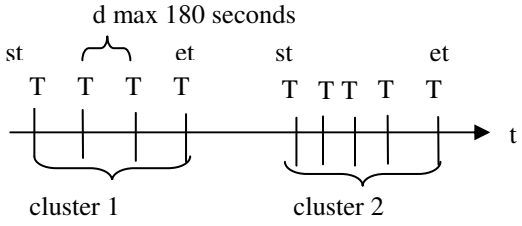
After the clusters are generated another algorithm (we call it *phrasing-algorithm*) analyses the cluster, if meta-phrase (definition/explanation, example, overview, repetition) occurs in each one. For this algorithm we also generated a specific thesaurus called *meta-phrase-thesaurus*. At the end of our procedure the user gets a result set with semantically enriched data of the searching word. And furthermore, a ranked list of the results. (see figure 2.)

5.1. Clustering

One of our goals is to be able to identify topics or a topic word in a stream of transcripts. The main steps of the clustering algorithm of the course works in the following way (see also Figure 3.):

1. take the query word / topic word of the user
2. take all patterns from the IT Thesaurus of the query word
3. take the lecture
4. build clusters so that the distance between two adjacent patterns is not more than 180 seconds, count the occurrence and set start time and end time of the cluster
5. next lecture and go to step 3

Figure 3. Example of the cluster algorithm



T = "topology" phrase from the thesaurus
st = start time of the cluster
et = end time of the cluster

In the table 1. there are some examples of these clusters. *Lecture_id* stands for a lecture from the course, *cn* is the cluster number, *word* is the topic word, *start* is the *start* time beginning inside the lecture, *end* is the end time and *wn* is the phrase/ word count inside this cluster. The word number *wn* for each cluster is the ranking value for the segment. (The higher the value *wn*, the higher the relevance of the segment.).

Table 1. Example set of clusters

lecture_id	cn	word	start	end	wn
1	1	topology	1600	2400	15
1	2	topology	2000	2200	4
2	1	topology	1020	1200	2
etc…					

5.2. Phrasing

Firstly we build a thesaurus for this so called "meta-phrase" and we store this thesaurus in a relational database. An example, definition/ explanation, repetition and overview in a course has a special pattern. For example for a repetition it is "letzte Stunde" (last lesson) or "wir sprachen über" (we spoke

about) and so on. We analyzed the course and built with this knowledge a thesaurus containing our four meta-phrases. We find for the four meta-phrases-thesaurus many of such patterns. The searching process of the phrasing algorithm works in the following way:

1. take a meta-phrase
2. take all patterns from the meta-phrase-thesaurus for this meta-phrase
3. search inside the start and the end time of the clusters for these patterns and store the result in the database
4. next cluster and go to step 3
5. next meta-phrase and go to step 2

6. Result and evaluation

The corpus of the videos consists of the course "Einführung in das WWW" held in the German language in the first semester 2005 at the Hasso Plattner Institut Potsdam. In the experiments we use only the speech of one speaker of the course. This part of the course consists of 16 lectures, each lecture being approximately 90 minutes in length. The video corpus has an over all length of approximately 1440 minutes and is stored as *RealMedia* files. The speech recognition software is trained with a standard tool in 15 Minutes and it is qualified with some special domain word (100 words) from the existing power point slides in 15 Minutes. So the training phase for the speech-recognition-software is approximately 30 Minutes long. As we mention in Chapter 3.1 the WER is between 20%-40% and the transcript of the course is stored in a relational database. We choose 17 words / topic words for our evaluation. The words in German are *Topologie, Bustopologie, Ringtopologie, Sterntopologie, TCP / IP, FDDI, Paketvermittlung, Token Passing, WLAN, MAC, Leitungsvermittlung, Verbindungsorientierte Dienst, Routing/Router, IP, OSI/Schichtenmodell, Ethernet / CSMA* and *DSL*

For our evaluation we use the recall and precision mass [8].

6.1. Cluster precision

We evaluate whether our generated clusters represent the theme of the searching word. For this, we take the cluster with the majority of the topic words *wn* and decide whether this cluster represented the theme accurately. And further more, if this is the best area in the whole course. Of course, that is a subject decision of our team which has to be evaluated in the future by the students' choice. The first ranked cluster of the 15 topic words matches exactly the proper start position of the theme. The other two areas don't match the topic. With this value we get a Precision of 88 % and of course the Recall of 88% too.

6.2. Meta-phrase recall and precision

Our first step is to analyze the occurrence of a definition/explanation, an example, an overview and repetition in each generated cluster. For this purpose we manually determine the frequency of the meta-phrase in each cluster. The table 2 show this analyst procedure. Where C is the cluster number, D/E is a definition/explanation, E is an example, O is an overview and R a repetition.

Table 2. Manually created results for each cluster

C	Meaning	D/E	E	O	R
1	topology	2	6	1	0
2	topology	4	0	2	0
3	topology	1	0	0	0
1	routing	2	2	1	0
etc…					

After this we evaluate the result-set supported from the cluster- and phrasing algorithm. We evaluate whether each hit in the result set, is correct (*y*), not correct (*n*), similar to the area of the topic (*T*) or a hit for another topic (*A*). Table 3 shows this exploration. Where *ph* stands for type of phrase, *E* stands for example, *E/D* stand for explanation/definition and *O* stand for overview. The sign *wn* stands for the occurrence of the words/ topic words inside the cluster. *c* stand for the cluster.

Table 3. Result-set from clustering and phrasing

automatic generated					manually			
clustering				phra -sing	validation			
C	meaning	time	wn	ph	y	n	T	A
1	topology	2000	15	E	X			
1	topology	1773	15	E			X	
1	topology	1729	15	E			X	
1	topology	1680	15	E/D	X			
etc…								

In the final result table 4 we sum up all **f**ound **h**its (Σ_{fo}) and the **c**orrect **h**its (Σ_{ch}) found by the algorithm. Σ_{ch} is the sum of all correct hits (*y*) and (T) the hits, which are similar to the topic. The column (Σ_{me}) stands for the **m**anually **e**valuated values. With these values we calculate the recall (rec.) and the precision (prec.) rate in percentage for each word. At the end of the table we sum up the values and calculate the total recall and the precision rate.

Table 4. Recall and precision

Word / Topic	Σfo	Σch	Σme	rec	*prec.*
topology	11	11	18	61	*100*
bus topology	7	5	9	78	*71*
ring topology	2	1	4	50	*50*
star topology	3	2	3	100	*67*
TCP / IP	22	18	27	81	*81*
FDDI	11	9	15	73	*81*
packet switching	22	15	28	78	*68*
Token Passing	7	6	8	87	*86*
WLAN	5	5	10	50	*100*
MAC	3	3	4	75	*100*
line switching	4	3	8	50	*75*
connection-oriented service	5	5	6	83	*100*
Routing/Router	21	16	25	84	*76*
IP	20	18	22	90	*90*
OSI	2	2	6	33	*100*
DSL	3	3	5	60	*100*
Ethernet / CSMA	4	4	7	57	*100*
Sum / Result	*152*	*126*	*204*	*76*	*82*

7. Conclusion and Outlook

In this paper, we present a system of intelligent retrieval systems that is specially designed for educational systems, and that allows the retrieval of documents or sections of documents from a multimedia knowledge base. We are able to confirm by experiments that the quality of several state-of-the-art speech recognition systems is good enough to be used in smart educational systems. The clustering and phrasing of the lecture were explained. The results show that it is possible to add meta data to the result set to support the students with helpful information. The precision is surprisingly high. It shows that is possible to get additional information with an easy algorithm. To improve the recall value we will conduct research to find better patterns for detecting the meta-phrase.

We are currently working on the improvement of our algorithm. The different sections inside a multimedia document must be detected fully automatically. User studies are also planned to evaluate the effectiveness of educational systems that are based on our architecture. Further more, our evaluation is based only on one speaker. In the future we will evaluate our results with more speakers.

We are also working on a "lecture-browser" for a simple navigation through the corpus of lectures, as we mentioned before in Chapter 3. This lecture-browser will help the students by their learning in an effective way. The combination of pedagogical and content description leads to novel forms of visualization and exploration of course lectures. With an existing prototype it is possible to navigate easily through the course material with a web-browser.

8. References

[1] Chen T., Meinel C., Schillings V., "Teleteaching Anywhere Solution Kit", *Hasso-Plattner-Institut für Softwaresystemtechnik GmbH*, Potsdam, http://www.tele-task.de (last access: 12/12/2005)

[2] Schillings V., Meinel C., tele-TASK - Teleteaching Anywhere Solution Kit. *Proceedings. ACM SIGUCCS* 2002, Providence (Rhode Island), USA, 2002, pp. 130-133

[3] Dongru Z., Yingying Z., "Video Browsing and Retrieval Based on Multimodal Integration", *Proceedings of the IEEE/WIC International Conference on Web Intelligence*. Halifax, Canada, 2003

[4] Hürst, W., "A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web", *IADIS International Journal on WWW/Internet*, Volume I, Number 1, 2003, pp. 43-58.

[5] Chau M., Jay F., Nunamaker Jr., Ming L., Chen H., "Segmentation of Lecture Videos Based on Text: A Method Combining Multiple Linguistic Features", *Proceedings of the 37th Hawaii International Conference on System Sciences*. Hawaii, USA, 2004.

[6] Carstensen K.-U., Ebert Ch., Endris C., Jekat S., Klabunde R., Langer H., *Computerlinguistik und Sprachtechnologie*, Spectrum Akademischer Verlag. Munich, Germany, 2004.

[7] Abowd, G.D., "Classroom 2000: An experiment with the instrumentation of a living educational environment", *IBM Systems Journal*, Vol. 38, No. 4, 1999, pp. 508-530.

[8] Baeza-Yates R., Ribeiro-Neto B., *Modern Information Retrieval,* Addison-Wesley, 1999, USA.

[9] Glass J., Hazen T.J., Hetherington L., Wang C., "Analysis and Processing of Lecture Audio Data: Preliminary Investigations." *Proceedings of the HLT-NAACL 2004 Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval,* Boston, MA, 2004, pp. 9-12.

[10] Herter E., Lörcher W., *Nachrichtentechnik,* Hanser. München, Wien, Germany, 1994, pp.44-45.

[11] Linckels S., Meinel Ch., Engel T., "Teaching in theCyber Age: Technologies, Experiments, and Realizations", *Proceedings of 3. Deutschen e-Learning Fachtagung der Gesellschaft für Informatik (DeLFI),* Rostock, Germany,2005, pp. 225 – 236