Question Answering from Lecture Videos based on an Automatic Semantic Annotation

Stephan Repp, Serge Linckels, Christoph Meinel
Hasso Plattner Institut (HPI), University of Potsdam
P.O. Box 900460, D-14440 Potsdam
{repp,linckels,meinel}@hpi.uni-potsdam.de

ABSTRACT

The number of digital lecture video recordings has increased dramatically. The accessibility, usability and the traceability of their content for students-use is limited. Therefore retrieval of audiovisual lecture recordings is a complex task. Speech recognition is applied to create a tentative and deficient transcription of the video recordings. The imperfect transcription is sufficient to generate semantic metadata serialized in an OWL file. A question answering system based on the automatically generated semantic annotations and a semantic search engine are presented. The annotation process is discussed, evaluated and compared to a perfectly annotated OWL file and, further, to a corrected transcript of the lecture.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process, Selection process; H.3.1 [**Content Analysis and Indexing**]: Indexing methods, Abstracting methods

General Terms

Algorithms, Experimentation, Management, Reliability

Keywords

multimedia retrieval, multimedia knowledge base, speech recognition, transliteration, ontology, semantic annotation, OWL, learning object

1. INTRODUCTION

The amount of educational content in electronic form is increasing rapidly. At the Hasso Plattner Institut (HPI) alone, 25 hours of university lecture videos about computer-science are produced every week. Most of them are published in the online Tele-TASK archive[1].

[1] `http://www.tele-task.de`

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ITiCSE '08, 2008, Madird, Spain.
Copyright 2008 ACM .....$5.00.

Although such resources are common, it is not easy for a user to find one that corresponds best to his/her expectations. This problem is mostly due to the fact that the content of such resources is often not available in machine readable form, i.e. described with metadata so that search engines, robots, or agents can process them. Indeed, the creation of semantic annotation neither is nor should be the task of the user or of the creator of the learning object. The user (e.g. a student) and the creator (e.g. a lecturer) are not necessarily computer-science experts who know how to create metadata in a specific formalism like XML, RDF or OWL. Furthermore, the creation of metadata is a subjective task and should be done with care. The automatic generation of reliable metadata is still a very difficult problem and currently a hot topic in the Semantic Web movement.

In this paper we will explore a solution to how to generate semantic annotations for university lectures. It is based on the extraction of metadata from two data sources — the content of the slides and the transliteration of an out-of-the-box speech recognition engine— and the mapping of natural language (NL) to concepts/roles in an ontology. The reliability of our solution is evaluated via different benchmark tests.

2. RELATED WORK

Using speech recognition to annotate videos is a widely used method [17, 5, 11]. Due to the fact that the slides carried most of the information, Repp et al. synchronized the imperfect transcript from the speech recognition engine automatically with the slide streams in post-processing [14].

Most approaches use out-of-the-box speech recognition engines, e.g. by extracting key phrases from spoken content [5].

In [6] a commercial speech recognition system is used to index recorded lectures. However, the accuracy of the speech recognition software is rather low; the recognition accuracy of the transliterations is approximately 22%-60%. It is also shown in [6] that audio retrieval can be performed with out-of-the-box speech-recognition-software. But little information can be found in literature about educational systems that use a semantic search engine for finding additional (semantic) information effectively in a knowledge base of recorded lectures.

A system for reasoning through multimedia e-Learning objects is described in [4]. An automatic speech recognition engine is used for keyword spotting. It extracts the taxonomic node that corresponds to the keyword, and associates it to the multimedia objects as metadata.

In proc. of ITICSE 2008, Madrid, Spain, p. 17 - 21

Two complete systems for recording, annotating, and retrieving multimedia documents are LectureLounge and MOM. LectureLounge [16] is a research platform and a system to automatically and non-invasively capture, analyze, annotate, index, archive and publish live presentations. MOM (Multimedia Ontology Manager) [3] is a complete system that allows the creation of multimedia ontologies, supports automatic annotation and the creation of extended text (and audio) commentaries of video sequences, and permits complex queries by reasoning through the ontology.

Based on the assertion that information retrieval in multimedia environments is actually a combination of search and browsing in most cases, a hypermedia navigation concept for lecture recordings is presented in [10].

An experiment is described in [7] where automatically extracted audio-visual features of a video were compared to manually annotations that were created by users.

3. EXTRACTION METHOD

3.1 Ontology Fundamentals

It has been realized that a digital library benefits from having its content understandable and available in a machine processable form, and it is widely agreed that ontologies will play a key role in providing a lot of the enabling infrastructure to achieve this goal. A fundamental part of our system is a common domain ontology. An existing ontology can be used or one can be built that is optimized for the knowledge sources.

An ontology is basically composed of a hierarchy of concepts (*taxonomy*) and a language. In the case of the first issue, we created a list of semantically relevant words from the domain of Internetworking, and organized them hierarchically. In the second case, we used *Description Logics* to formalize the semantic annotations.

Description Logics (DL) [1] are a family of knowledge representation formalisms that allow the knowledge of an application domain to be represented in a structured way and to reason about this knowledge. In DL, the conceptual knowledge of an application domain is represented in terms of *concepts* (unary predicates) such as IPAddress, and *roles* (binary predicates) such as ∃composedOf. Concepts denote sets of individuals and roles denote binary relations between individuals. Complex descriptions are built inductively using concept constructors which rely on basic concepts and role names. Concept descriptions are used to specify terminologies that define the intentional knowledge of an application domain. Terminologies are composed of *inclusion assertions* and *definitions*. The first impose necessary conditions for an individual to belong to a concept. E.g. to impose that a router is a network component that uses at least one IP address, one can use the inclusion assertion: $\mathsf{Router} \sqsubseteq \mathsf{NetComp} \sqcap \exists \mathsf{uses.IPAddress}$. Definitions allow us to give meaningful names to concept descriptions such as $\mathsf{LO}_1 \equiv \mathsf{IPAdress} \sqcap \exists \mathsf{composedOf.HostID}$.

The semantic annotation of five learning objects is shown in figure 1, describing the following content:

Figure 1: Example of terminology concerning learning objects.

$\mathsf{LO}_1 \equiv \mathsf{IPAddress}$
$\mathsf{LO}_2 \equiv \mathsf{TCPIP} \sqcap \exists \mathsf{uses.IPAddress}$
$\mathsf{LO}_3 \equiv \mathsf{IPAddress} \sqcap \exists \mathsf{composedOf.HostID}$
$\mathsf{LO}_4 \equiv \mathsf{IPAddress} \sqcap \exists \mathsf{composedOf.NetworkID}$

- LO_1: general explanation about IP addresses,
- LO_2: explanation that IP addresses are used in the protocol TCP/IP,
- LO_3: explanation that an IP-address is composed of a host identifier,
- LO_4: explanation that an IP-address is composed of a network identifier,

Some advantages of using DL are the following: firstly, DL terminologies can be serialized as OWL (*Semantic Web Ontology Language*) [15], a machine-readable and standardized format for semantically annotating resources (see section 3.5). Secondly, DL allow the definition of detailed semantic descriptions about resources (i.e. restrictions of properties), and logical inference from these descriptions [1].

3.2 Natural Language Processing

The way our NL processing works is described in detail in [8]. To make this paper self-containing, we will briefly summarize the major ideas.

The system masters a domain dictionary L_H over an alphabet Σ^* so that $L_H \subseteq \Sigma^*$. The semantics are given to each word by classification in a hierarchical way (taxonomy) . This means, for example, that words such as "IP-address", "IP adresse" and "IP-Adresse" refer to the concept IPAddress in the taxonomy. The mapping function φ is used for the semantic interpretation of a NL word $w \in \Sigma^*$ so that $\varphi(w)$ returns a set of valid interpretations, e.g. φ("IP Addresse") ={IPAddress}.

The system allows a certain tolerance regarding spelling errors, e.g. the word "comXmon" will be considered as "common", and not as "uncommon". Both words "common" and "uncommon" will be considered for the mapping of "comXXmon". In that case the mapping function will return two possible interpretations, so that:

$$\varphi(\text{"comXXon"}) = \{\mathsf{common}, \mathsf{uncommon}\}.$$

A dictionary of synonyms is used. It contains all relevant words for the domain — in our case: networks in computer-science — and at least all the words used by the lecturer (audio data) and in the slides.

3.3 Identification of Relevant Keywords

Normally, lectures have a length of around +/- 90 minutes, which is much too long for a simple learning object. If a student is searching for particular and precise information, (s)he might not be satisfied if a search engine yields a complete lecture. Therefore, we split such lectures in shorter learning objects. In the current state of our solution, this pre-processing is done manually.

For us, a learning object is composed of two data sources: the audio data and the content of the slides. In the case of the first issue, the audio data is analyzed with an out-of-the-box speech recognition engine. After a normalization

pre-processing — i.e. deleting stop-words and stemming of the words — the stems are stored in a database. This part of our system has already been described in [13, 14].

Formally, the analysis of a data source is done with the function μ that returns a set of relevant words in their canonical form, written:

$$\mu(\mathsf{LO}_{source}) = \{w_i \in L_H, i \in [0..n]\} \backslash S$$

where *source* is the input source with *source* $\in$ {audio only, slides only, audio and slides}, and S is the set of stop words, e.g. S ={"the", "a", "hello", "thus"}.

3.4 Ranking of Relevant Concepts and Roles

Independent of the data source used (audio only, slides only, audio and slides), the generation of the metadata always works the same way. The relevant keywords from the data source identified by the function μ are mapped to ontology concepts/roles with the function φ as explained in section 3.2.

It is not useful to map all identified words to ontology concepts/roles because this would create to much overload. Instead, we focus on the most pertinent metadata for the particular learning object. Thus we implemented a simple ranking algorithm.

The algorithm works as follows: We compute for each identified concept/rule its hit-rate h, i.e. its frequency of occurrence inside the leaning object. Only the concepts/roles with the maximum (or d^{th} maximum) hit-rate compared to the hit-rate in the other learning objects are used as metadata. E.g. the concept Topology has the following hit-rate for the five learning objects (LO_1 to LO_5):

	LO_1	LO_2	LO_3	LO_4	LO_5
h	0	4	3	7	2

This means that the concept Topology was not mentioned in LO_1 but 4 times in LO_2, 3 times in LO_3 etc.

We now introduce the rank d of the learning object i.e. the hit-rate of a concept. For a given rank, e.g. $d = 1$, the concept Topology is relevant only in the learning object LO_4 because it has the highest hit-rate. For $d = 2$ the concept is associated to the learning objects LO_4 and LO_2, i.e. the two learning objects with the highest hit-rate.

3.5 Semantic Annotation Generation

The semantic annotation of a given learning object is the conjunction of the mappings of each relevant word in the source data written:

$$\mathsf{LO} = \prod_{i=1}^{m} rank_d \ \varphi(w_i \in \mu(\mathsf{LO}_{source}))$$

where m is the number of relevant words in the data source and d the rank of the mapped concept/role. The result of this process is a valid DL description similar to that shown in figure 1. In the current state of the algorithm we do not consider complex role, e.g. $\exists R.(A \sqcap \exists S.(B \sqcap A))$, where A, B are basic concepts and R, S are roles. We also try to use a very simple DL, e.g. negations $\neg A$ are not considered.

One of the advantages of using DL is that it can be serialized in a machine readable form without losing any of its details. Logical inference is possible when using these annotations. The example shows the OWL serialization for the following DL-concept description:

$$\begin{aligned} \mathsf{LO}_1 \equiv \ & \mathsf{IPAddress} \sqcap \\ & \exists \mathsf{isComposedOf}.(\mathsf{Host\text{-}ID} \sqcap \mathsf{Network\text{-}ID}) \end{aligned}$$

defining a concept name (LO_1) for the concept description saying that an IP address is composed of a host identifier and a network identifier.

4. EVALUATION AND DISCUSSION

4.1 Preliminaries

4.1.1 Search Engines

The standard keyword-based search engine works in a classical way by browsing the textual content of the resource. If one (relevant) keyword is found then the resource is considered as being a hit, i.e. a relevant document. The search engine is optimized in so that it does not consider stop words. The semantic search engine is described in detail in [9]. It reviews over the OWL-DL metadata and computes how much the description matches the query. In more detail, it quantifies the semantic difference between the query and the DL-concept description.

4.1.2 Knowledge Source

We used the online tele-TASK archive[2] that contains hundreds of recorded university lectures as a knowledge base. We selected the lecture on Internetworking (with a total of 100 minutes of lectures), which we split into 40 smaller units, i.e. multimedia learning objects. Each learning object has a duration of approximately 2.5 minutes. The audio data of the lecturer, i.e. the speech of the lecturer, were transcoded with an out-of-the box speech recognition software. The software was trained for +/- 15 minutes. Furthermore, some domain specific words were added to its dictionary. All together, we spent some 30 minutes to prepare and train the speech recognition software. For the transliteration, a word-accuracy of approximately 60% is measured. The stemming as described in section 3.3 was done with the "Porter stemmer" [12].

4.1.3 Evaluation Criteria

A set of 107 NL questions on the topic Internetworking was created. We worked out questions that students might ask, e.g. "*What is an IP-address composed of?*", "*How does a data packet find its way through a network?*", etc. For each question we also indicated the relevant answer that should be delivered by the search engine. We call an answer from a search engine a *perfect hit* (PH) if it yields only and exactly the relevant answer with no supplementary information. Our evaluation refers to the common *recall* (R) and *precision* (P) factor [2]. The recall R_1 (R_5 or R_{10}) considers only the first (first five or ten) hit(s) of the result set.

4.2 Results

The evaluation is based on four different semantic sources. Each was serialized as an OWL file. The different semantic descriptions were generated based on the following data:

- manually generated (M)
- generated from the slides (S)

[2] `http://www.tele-task.de/`

- generated from the imperfect transcript of the speech recognition of the audio data (T)
- generated from the perfect transcript of the audio data (PT).

Four different test scenarios were elaborated to evaluate the quality of the retrieval from the different semantic sources M, S, T, PT, as well as combinations of some of those. Here is an overview of the test settings:

1. a random retrieval: for each query the system simply yields 6 random results called "rand6".
2. a standard keyword-based search in the semantic sources S, T and PT.
3. a semantic search engine (as described in [9]) in the semantic sources M, S, T and PT.
4. a combination of different semantic sources based on the ranking criteria (see section 3.4). The configurations are the following:

$$[< source >]_{\text{ranking}}$$

where $< source >$ stands for the data source (M, S, T, or PT), and $< ranking >$ stands for the ranking ration (0 is no ranking at all, all concepts are selected, i.e. $d = 0$, and r ranking with $d = 2$). E.g. $[T+S]_2$ means that the metadata from the transcript (T) and from the slides (S) are combined (set union) and that the result is ranked with $d = 2$.

4.2.1 Outcomes concerning the search engine

	PH	R	P
rand6	0	16	16
T	12	77	6
PT	13	80	6
S	7	83	12

Table 1: The result (in %) of the random retrieval and the keyword-based search.

The evaluation of the random retrieval (1) and the keyword-based search (2) is shown in table 1. The results of the random retrieval (rand6) are not surprising. Perfect hits (PH) are statistically impossible and the recall (R) is bad. As for the keyword-based search, the values for R and P are also bad for all semantic sources (T, PT and S). The recall is high with a low precision. It is evident that these two configurations are not acceptable for use in any question-answering system.

The results of the evaluation of the semantic search engine (3) and different combinations (4) are shown in table 2. First, let us analyze the results of the uncombined semantic descriptions (first 7 lines). The following statements can be made: it is not surprising that the best search results were achieved with the manually generated semantic description (M), with 76% of perfect hits and 91% of recall. However, the semantic description from the perfectly transcribed lecture video (PT) is only slightly better than the automatically generated one (T). For T_0 the recall is 80% and for PT_0 it is

	PH	R_1	R_5	R_{10}	R	P
M	76	87	89	91	91	52
$[S]_0$	11	27	59	73	86	9
$[T]_0$	16	27	50	65	80	6
$[PT]_0$	16	23	51	66	84	9
$[S]_2$	16	26	57	74	79	11
$[T]_2$	14	29	48	60	69	8
$[PT]_2$	11	32	55	63	74	8
$[T + S]_0$	16	31	50	68	88	6
$[T + S]_2$	11	37	58	68	78	8
$[PT + S]_0$	13	30	49	71	88	6
$[PT + S]_2$	13	35	63	70	80	8

Table 2: Result (in %) of the semantic search.

84% with a precision of 6 and 9 respectively. The perfect hits are identical for both. A similar observation can be made by comparing T_2 with PT_2. An interesting outcome is that the semantic description from the slides (S) carries most of the information compared to T and PT. This is true for both rankings $d = 0$ and $d = 2$. A possible explanation might be that most of the important subjects are summarized on the lecture's slides and that therefore this knowledge source is semantically sufficient. Another explanation might be that — at least in our case — the lecturer mentioned other different concepts that were not directly related to the topic of the slides. These concepts have, however, been detected by the transcoding process.

Let us now analyze the results of the combined semantic descriptions (last 4 lines in table 2). The most striking outcome of our evaluation is that the combination of slide data and audio data improves the quality of the semantic description and thus the quality of the yielded results by a search engine. $[T + S]_2$ and $[PT + S]_2$ have much greater recall ($R_1 = 37\%$ and $R_1 = 35\%$ respectively, and $R_5 = 58$ and $R_5 = 63$ respectively) than without a combination of both semantic sources (see figure 2). A plausible explanation might be that the lecturer does not necessarily explain only the content of the slides but also speaks further and provides supplementary information that is relevant. This outcome should also be evidence for all students that attending on campus classes is still more complete than only studying the slides from the lecture.

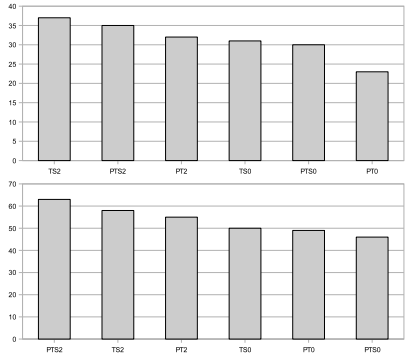

Figure 2: (1) the recall R_1 and (2) the recall R_5

4.2.2 Outcomes concerning the learner

As stated in the introduction, the aim of our research work is to give the user the technological means to quickly find the pertinent information. For the lecturer or the system administrator, the aim is to minimize the supplementary work a lecture may require in terms of post-production, e.g. creating the semantic description.

Let us focus in this section on the completely automated generation for semantic descriptions (T, S and its combination [T + S]). In such a configuration with a fully automated system $[T + S]_2$, a learner's question will be answered correctly in 37% of the cases by reading only the first result, and in 58% of the cases if the learner considers the first five results that were yielded. This score can be raised by using an improved speech-recognition engine or by manually reviewing and correcting the transcripts of the audio data. In that case $[T + S]_2$ allows a recall of 63% (70%) while reading the first 5 (10) returned results.

In practice, 63%(70%) means that the learner has to read at most 5 (10) learning objects before he finds the pertinent answer to his question. Let us recall that a learning object has an average duration of 2.5 minutes, so that the learner must spent — in the worst case — $5 * 2.5 = 12.5$ minutes (25 minutes) before (s)he gets the required answer.

5. CONCLUSION

In this paper we have presented an algorithm for generating a semantic annotation for university lectures. It is based on three input sources: the textual content of the slides, the imperfect transliteration and the perfect transliteration of the audio data of the lecturer. Our algorithm maps semantically relevant words from the sources to ontology concepts and roles. The metadata is serialized in a machine readable format, i.e. OWL. We have shown that the metadata generated in this way can be used by a semantic search engine.

Although the quality of the manually generate metadata is still better than the automatically generated ones, it is sufficient for use as a reliable semantic description in question-answering systems. In the worst case, a learner has to read up to 5 short learning objects to get the pertinent answer.

6. REFERENCES

[1] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[2] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[3] M. Bertini, A. D. Bimbo, C. Torniai, R. Cucchiara, and C. Grana. Mom: Multimedia ontology manager. a framework for automatic annotation and semantic retrieval of video sequences. In *ACM SIGMM*, pages 787–788, 2006.

[4] M. Engelhardt, A. Hildebrand, D. Lange, and T. C. Schmidt. Reasoning about eLearning Multimedia Objects. In *International Workshop on Semantic Web Annotations for Multimedia (SWAMM)*, 2006.

[5] A. Haubold and J. R. Kender. Augmented segmentation and visualization for presentation videos, 2005.

[6] W. Hürst, T. Kreuzer, and M. Wiesenhütter. A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web. In *IADIS Internatinal Conference WWW/Internet (ICWI)*, pages 135–143, 2002.

[7] A. Jaimes, T. Nagamine, J. Liu, K. Omura, and N. Sebe. Affective meeting video analysis. In *IEEE Multimedia and Expo*, pages 1412–1415, 2005.

[8] S. Linckels and C. Meinel. Resolving ambiguities in the semantic interpretation of natural language questions. In *Intelligent Data Engineering and Automated Learning (IDEAL)*, volume 4224 of *LNCS*, pages 612–619, 2006.

[9] S. Linckels, H. Sack, and C. Meinel. Optimizing the retrieval of pertinent answers for nl questions with the e-librarian service. In *Service Matchmaking and Resource Retrieval in the Semantic Web (SMR2)*, 2007.

[10] R. Mertens, H. Schneider, O. Müller, and O. Vornberger. Hypermedia navigation concepts for lecture recordings. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, pages 2480–2847, 2004.

[11] C.-W. Ngo, F. Wang, and T.-C. Pong. Structuring lecture videos for distance learning applications. In *Multimedia Software Engineering*, pages 215– 222, 2003.

[12] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[13] S. Repp, S. Linckels, and C. Meinel. Towards to an automatic semantic annotation for multimedia learning objects. In *Proceedings of the International Workshop on Educational Multimedia and Multimedia Education 2007, Augsburg, Bavaria, Germany, September 28, 2007*, pages 19–26. ACM, 2007.

[14] S. Repp, J. Waitelonis, H. Sack, and C. Meinel. Segmentation and annotation of audiovisual recordings based on automated speech recognition. In *Intelligent Data Engineering and Automated Learning - IDEAL 2007, Birmingham, UK, December 16-19*, volume 4881 of *LNCS*, pages 620–629. Springer, 2007.

[15] W. W. W. C. W3C. *OWL Web Ontology Language*. `http://www.w3.org/TR/owl-features/`, 2004.

[16] P. Wolf, W. Putz, A. Stewart, A. Steinmetz, M. Hemmje, and E. Neuhold. Lecturelounge – experience education beyond the borders of the classroom. *International Journal on Digital Libraries*, 4(1):39–41, 2004.

[17] N. Yamamoto, J. Ogata, and Y. Ariki. Topic segmentation and retrieval system for lecture videos based on spontaneous speech recognition. In *European Conference on Speech Communication and Technology*, pages 961–964, 2003.