Automatische Erzeugung Semantischer Annotationen für Vorlesungsvideos

Stephan Repp[1], Serge Linckels[1,2], Christoph Meinel[1]

[1]Hasso Plattner Institut (HPI), Universität Potsdam
Prof.-Dr.-Helmert-Str. 2-3, D-14482 Potsdam
[2]Luxembourg International Advanced Studies
in Information Technologies (LIASIT)
6, rue R. Coudenhove-Kalergi, L-1359 Luxembourg
{repp, linckels, meinel}@hpi.uni-potsdam.de

Abstract: Die Anzahl der aufgezeichneten digitalen Vorlesungen nimmt rapide zu. Immer mehr Hochschuleinrichtungen nutzen die Möglichkeit, ihre Vorlesungen als Videos aufzuzeichnen, in multimedialen Datenbanken abzulegen und für den Zugriff über das Internet bereitzustellen. Bislang ist die detaillierte Suche in diesen Multimedia Daten jedoch nur begrenzt möglich. Der Hauptgrund hierfür ist in dem Umstand zu suchen, dass eine manuelle semantische Annotation aus Kostengründen ausscheidet - ein gutes automatisiertes Verfahren bislang jedoch nicht zur Verfügung steht. Die Entwicklung einer solchen automatischen semantischen Annotation stellt daher insbesondere für den Bereich des E-Learnings ein Desiderat und eine anspruchsvolle Aufgabe für die Informatik dar. Der vorliegende Aufsatz stellt eine Methode für die automatische semantische Annotation dar, die aus den Transkripten des Vortragenden erzeugt wird. Die fehlerbehafteten Texte eines Spracherkenners reichen hierbei aus, um eine semantische Annotation der Vorlesungen durchzuführen. Die Annotationen werden in einer OWL-Datei gespeichert und einer semantischen Suchmaschine zugeführt. Ein Frage-Antwort-System, basierend auf den automatisch generierten Annotationen, wird vorgestellt, evaluiert und zur Diskussion gestellt.

1 Einführung

Viele Universitäten und andere Hochschulen nutzen die Möglichkeit ihre Veranstaltungen aufzuzeichnen und diese multimedialen Daten einem breiten Interessentenkreis zugänglich zu machen. Die Summe der aufgezeichneten Vorlesungen steigt daher täglich rapide an. Alleine am Hasso Plattner Institut (HPI) werden jede Woche 25 Stunden Vorlesung der Informatik aufgezeichnet. Diese Vorlesungen werden online[1] veröffentlicht. Die Suche in diesen Archiven und der Zugriff auf eine Videosequenz stellen die Bildungseinrichtungen vor verschiedene Probleme: So ist die manuelle Annotation der Multimedia-Daten allein aus Kostengründen in der Regel nur für einzelne, nicht aber für die Vielzahl der vorhandenen Videos möglich. Doch selbst die manuelle Annotation einzelner Videos oder Videoabschnitte erweist sich als nicht unproblematisch. Es handelt sich hierbei um eine

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

In Proc. DeLFI2008, pp. 149 - 160, ISBN: 978-3-88579-226-0

ausgesprochen eintönige Arbeit, die jedoch sehr sorgfältig ausgeführt werden muss, da die Erstellung von Annotationen sonst zu subjektiven Ergebnissen führen kann. Sie wird in der Regel von Studenten oder Dozenten ausgeführt, die üblicherweise keine Experten in der Erstellung von Annotationen in einem speziellen Format wie XML, RDF oder OWL sind. Eine automatische Annotation ist daher dringend erforderlich. Das Format der Annotationen sollte maschinenlesbar sein, damit die gewonnenen Daten von Suchmaschinen, Robots oder Agenten entsprechend verarbeitet werden können.

Die Erstellung von automatischen Annotationen stellt somit eine neue Herausforderung und ein spannendes Thema im Bereich des Semantic Web dar. In diesem Beitrag zeigen wir eine Möglichkeit auf, wie die semantische Annotation von Vorlesungsaufzeichnungen automatisch erstellt werden kann. Diese Lösung basiert auf der Extraktion von Meta-Daten von zwei Datenquellen: dem Inhalt der Präsentationsfolien und den Sprachtranskripten eines Spracherkenners (out-of-the-box System). Die natürliche Sprache des Sprechers wird den entsprechenden Konzepten und den entsprechenden Rollen einer Ontologie zugeordnet. Im Vergleich mit einer manuell erstellten Annotation evaluieren wir unsere Arbeit und stellen unsere Ergebnisse zur weiteren Diskussion.

2 Relevante Arbeiten

Die Verwendung von Sprachtranskripten und die Nutzung von Vortragsfolien für die Annotation von Vorlesungen sind zwei weit verbreitete Methoden [HK05, NWP03, RM06a, RM06b, YOA03]. Es hat sich jedoch gezeigt, dass Vortragsfolien bei der Stichwortsuche die besten Resultate liefern. Werden während der Vorlesung keine log-files über die Folienwechsel angelegt, gehen die zeitlichen Informationen der Folienübergänge verloren; sie können jedoch aus den Folieninhalten und dem Sprachtranskript in einem Postprozess berechnet werden [RWSM07]. Die bislang durchgeführten Arbeiten nutzen die Ergebnisse eines Spracherkenners vorrangig, um reine Schlüsselwörter aus den Vorlesungen zu extrahieren [HK05]. So verwendet beispielsweise Hürst et. al. [HKW02] einen kommerziellen Spracherkenner um die Vorlesungsvideos zu annotieren. Die Erkennungsqualität der Spracherkenner ist jedoch so gering, dass nur 22%-60% korrekt in einen Wortstrom übersetzt werden. Zugleich zeigt Hürsts Methode jedoch auch, dass ein Retrieval mit diesen Audiodaten durchgeführt werden kann. Der Grund hierfür liegt in der hohen Redundanz der Sprache. So werden Schlüsselwörter in der Regel häufiger als einmal in der Vorlesung genannt. Ein falsch erkanntes Schlüsselwort hat in der Regel nur geringe Auswirkung auf das Suchergebnis. Diese Methode zur Schlüsselwortsuche, wie sie von Hürst u.a. vertreten wird, kann durch Vorgehensweisen ergänzt werden, wie sie zum Beispiel von [SW06] implementiert worden sind. Ihnen ist gemeinsam, dass sie die Benutzer von Archiven aktiv in den Annotations-Prozess involvieren. Durch ein so genanntes *tagging* werden die Videos mit Bemerkungen der Benutzer angereichert. Der Anfragende hat nun die Möglichkeit, in diesen Bemerkungen zu suchen.

Ein System, das Schlussfolgerungen aus den multimedialen Lernobjekten ziehen kann, ist in [EHLS06] beschrieben. Die Veröffentlichung zeigt auf, wie Transkripte des Spracherkenners genutzt werden können, um Schlüsselwörter aus den Vorlesungen zu generieren.

Diese extrahierten Schlüsselwörter werden einem entsprechenden Taxonomie-Knoten zugeordnet und ergänzen so das Multimedia-Objekt.

Ein System für die Aufzeichnung und Annotation von Multimedia-Dokumenten sowie ein System zur Suche in diesen Datenbeständen sind LectureLounge und MOM. LectureLounge [WPS+04] ist eine Forschungsplattform und ein System zur automatischen Analyse, Annotation, Indexierung, Archivierung und Veröffentlichung von Präsentationen. Multimedia Ontology Manager [BBT+06] ist ein System zur Erstellung von Multimedia Ontologien, zur Unterstützung von Annotationen und für die Erzeugung von ergänzenden Texten und Audiokommentaren zu einzelnen Videosequenzen; es erlaubt außerdem die Beantwortung komplexer Anfragen durch Schlussfolgerungen. Basierend auf der Aussage, dass Informationssuche eine Kombination aus Bereitstellung, Filterung und Ranking der Dokumente einerseits und aus aktivem Browsing der Ergebnisse andererseits darstellt, präsentiert [MSMV04] ein Hypermedia-Navigation-Konzept.

Die Entwicklung von semantischen Suchmaschinen ist im Unterschied zur Entwicklung von Suchmaschinen zur Schlüsselwortsuche bislang wenig ausgereift. So gibt es bisher nur wenige Veröffentlichungen über semantische Suchmaschinen, die automatisch generierte, semantische Annotation nutzen. In [CFRN05] wird ein Frage/Antwort System für vorsegmentierte Vorlesungsvideos vorgestellt. Über eine trainierte Mustererkennung werden die verschiedenen Muster der Antworten den entsprechenden Fragen zugeordnet; dabei werden die Ergebnisse mit einer natürlichen Sprachverarbeitungsmethode verglichen.

Die vorliegende Veröffentlichung basiert auf den Arbeiten [RLM07, LM06]. Zusätzlich zu [RLM07] werden die Lernobjekte nicht manuell vorbestimmt (Segmentierung des Videos), sondern durch die Folienübergänge automatisch festgelegt. Darüber hinaus wird die zeitliche Abfolge der Lernobjekte mit berücksichtigt und die Ergebnisse mit dem Evaluationsmaß MRR ausgewertet.

3 Frage-Antwort-System

Das Kapitel führt in die Theorie der Ontologien und der natürlichen Sprachverarbeitung ein (Abschnitt 3.1). Es beschreibt in einem zweiten und dritten Schritt die Anfragebearbeitung (Abschnitt 3.2) und Methoden der Extraktion semantischer Komponenten, Rollen und Konzepte aus Vorlesungstranskripten und Folienströmen (Abschnitt 3.3) und erläutert schließlich, wie die Informationen in einer maschinenlesbaren Form für die semantische Suchmaschine bereit gestellt werden können (Abschnitt 3.4).

3.1 Grundlagen

Bei der Annotation von Inhalten kommt den Ontologien eine Schlüsselrolle zu, da sie die Beschreibung von “Wissen” aus den semantischen Zusammenhängen ermöglichen. Ein fundamentaler Teil unseres Systems ist eine gemeinsame Ontologie, die auf das Themengebiet (in unserem Fall Internetworking) angepasst ist. Bestandteile einer Ontolo-

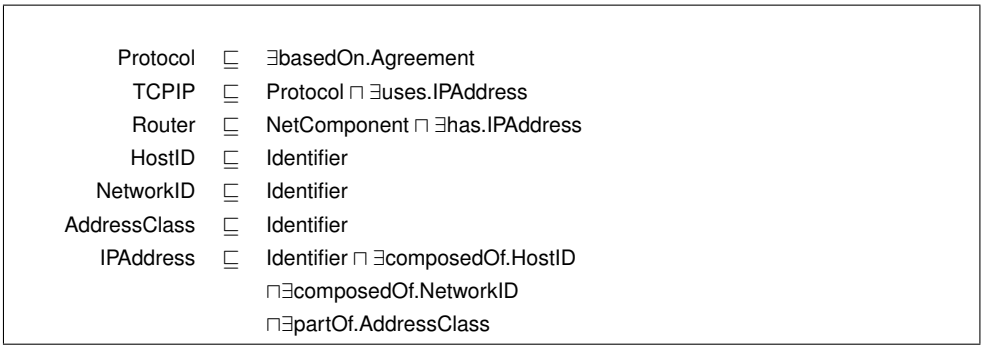

Abbildung 1: Beispiel einer Netzwerk Terminologie.

gie sind eine Hierarchie von Konzepten (*taxonomy*) und eine Sprache. Das Konzept besteht hierbei aus einer semantisch geordneten hierarchischen Struktur. Die Sprache wird durch die Deskriptive Logik (DL) abgebildet, die die Beziehungen der Konzepte untereinander beschreibt. Deskriptive Logik [BCM$^+$03] ist ein Formalismus der Wissensrepräsentation. Er ermöglicht, dass "Wissen" strukturiert, maschinenlesbar und einheitlich abgelegt werden kann. Mit Hilfe dieses Formalismus können nun Schlussfolgerungen über das "Wissen" berechnet werden. In der DL wird das konzeptionelle "Wissen"einer Domain mit sogenannten Konzepten repräsentiert, wie zum Beispiel IPAddress. Die Beziehungen der Konzepte untereinander werden mit Rollen ausgedrückt, wie zum Beispiel ∃composedOf. Komplexe semantische Beschreibungen können nun aus den Basiskonzepten und einigen Rollen zusammengestellt werden. Beispiele der Notationen für eine Konzept-Beschreibung mit Hilfe von DL sind:

- top-concept ($\top$) und bottom-concept ($\bot$). Sie bezeichnen alle Individuen und das leere Datenset;
- UND Verknüpfung (conjunction) ($\sqcap$);
- existentielle Restriktion (existential restriction) ($\exists r.C$), z.B.: IPAddress. $\sqcap\exists$composedOf.HostID bedeutet, dass eine IP Adresse aus einer Host-ID besteht.

Konzept-Beschreibungen (Terminologien) werden genutzt, um explizites Wissen in einer Domain zu beschreiben. Eine Terminologie besteht aus *inclusion assertions* und *definitions*. *Inclusion assertions* beschreiben notwendige Bedingungen für die Individuen um ein Konzept hinreichend darzustellen. Angenommen, ein Router ist ein Netzwerkelement und dieser Router verwendet mindestens eine IP Adresse, so lautet die Beschreibung der *inclusion assertions* in DL: Router $\sqsubseteq$ NetComp $\sqcap$ ∃uses.IPAddress. Definitionen werden genutzt um aussagekräftige Namen der Konzeptbeschreibungen zu vergeben: $\mathsf{LO}_1 \equiv$ IPAdress $\sqcap$ ∃composedOf.HostID.

Die Abbildung 1 zeigt ein Beispiel einer Nezwerk Terminologie. Die semantische Annotation von fünf Lernobjekten wird in der Abbildung 2 dargestellt. Die vier Lernobjekte

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

Abbildung 2: Beispiel für die Terminologie einiger Lernobjekte.

beschreiben folgenden Inhalt:

LO_1: Allgemeine Erklärung über IP Adressen
LO_2: Eine IP-Adressen wird vom Protokoll TCP/IP benutzt
LO_3: Eine IP-Adresse besteht aus einem Host-Identifier
LO_4: Eine IP-Adresse besteht aus einem Network-Identifier
Zusammengefasst lassen sich folgende Vorteile der DL benennen:

- Erstens: DL Terminologien können als OWL-Dateien serialisiert werden (*Semantic Web Ontology Language*) [W3C04], die maschinenlesbar und ein Standard sind.

- Zweitens: DL erlaubt es, detaillierte semantische Beschreibungen von Ressourcen anzufertigen. Mit Hilfe dieser Beschreibungen können logische Schlussfolgerungen gezogen und neue Zusammenhänge aus den Daten erschlossen werden [BCM$^+$03].

- Drittens: Zwischen der DL und der natürlichen Sprache (NL) besteht ein enger Zusammenhang. Dies ist bei einer Anfrage in natürlicher Sprache von Vorteil [Sch93].

3.2 Anfragebearbeitung

Die vorgestellten theoretischen Erläuterungen bilden die Grundlage für die Entwicklung der im Folgenden dargestellten Vorgehensweise bei der Anfragebearbeitung.

Das System besteht aus einem Domain Lexikon L_H mit einem Alphabet Σ^*, sodass $L_H \subseteq \Sigma^*$ ist. Die Semantik wird durch die Einordnung jedes Wortes in die Hierarchie bzw. in die Taxonomie erreicht. Das bedeutet z.B., dass Wörter wie “IP-address”, “IP adresse” und “IP-Adresse” in der Taxonomie dem Konzept IPAddress zugeordnet werden. Die Zuordnungs-Funktion φ benutzt die semantische Interpretation eines NL Wortes $w \in \Sigma^*$, sodass $\varphi(w)$ eine Menge von gültigen Interpretationen liefert. Zum Beispiel:

$$\varphi(\text{”IP Addresse”}) = \{\mathsf{IPAddress}\}.$$

Diese Funktion wird als Zuordnung in einer Datenbank abgelegt.

Zusätzlich wird ein Synonym-Lexikon benutzt. Es enthält alle relevanten Wörter für diese Domain — in unserem Fall Internetworking— welche in den Folien und in der Vorlesung durch den Vortragenden vewendet werden.

3.3 Extraktion der relevanten Konzepte und Rollen

Für die Suche nach einem bestimmten, genau definierten Thema, ist die normale Vorlesungslänge (ungefähr 90 Minuten) eines Multimedia-Learnobjektes zu lang. Aus diesem Grund spalten wir unsere Vorlesung mit Hilfe der Folienübergänge auf. Spricht der Vortragende über eine Folie, so stellt dieser Zeitraum nun ein multimediales Lernobjekt dar. Damit die Zeiten der Folienübergänge erhalten bleiben, muss eine Log-Datei während der Präsentation erstellt werden. Eine andere Möglichkeit besteht darin, die Zeiten in einem Post-Prozess zu berechnen [RWSM07]. Für unsere Versuche sind die Folienumbrüche manuell erstellt worden, um die entstehenden Fehler durch eine Log-Datei oder den Post-Prozesses auszuschließen.

Ein Lernobjekt besteht aus zwei Datenquellen: der Sprache des Vortragenden und dem Inhalt der Folien. Das Sprachsignal wird mit Hilfe eines Spracherkenners in ein Transkript (Textstrom des gesprochenen Wortes des Vortragenden) umgewandelt. Nach einer standardisierten Vorverarbeitung (Löschen von Stoppwörtern, Stemming [Por80] der Wörter) - werden die Wortstämme mit den entsprechenden Zeitmarken in einer Datenbank abgelegt.

Die Daten der jeweiligen Quelle werden nach folgender Funktion analysiert: μ liefert ein Datenset von relevanten Wörtern in der Form:

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

$source$ bezeichnet die vewendete Datenquelle; $source \in$ {Sprachtranskript, Folien und Kombination Folien / Sprachtranskript}; S ist die Stoppwortliste, z.B.: $S =$ {“der”, “a”, “so”, “und”}.

3.4 Zuordnung der Konzepte / Rollen zu den Lernobjekten (Ranking)

Die Erzeugung der Annotation erfolgt unabhängig von den Datenquellen in der gleichen Weise. Das relevante Schlüsselwort, das durch die Funktion μ erkannt wurde, wird einem Konzept oder einer Rolle mit der Funktion φ zugeordnet.

Diese Konzepte und Rollen treten in den einzelnen Datenquellen (Folien, Sprache) auf und können somit den einzelnen Lernobjekten zugeordnet werden. Damit der Fokus auf die Extraktion der wichtigsten Konzepte eines Lernobjektes erfolgt, arbeitet der Zuordnungs-Algorithmus wie folgt: Für jedes identifizierte Konzept berechnen wir die Auftrittshäufigkeit (Auftrittsfrequenz) h, die Frequenz des Auftretens des Konzeptes innerhalb des Lernobjektes. Nur die Konzepte mit der maximalen Auftrittsfrequenz (oder d-ten Maximum) verglichen mit den Auftrittsfrequenzen in den anderen Lernobjekten werden für die Annotation verwendet. Zum Beispiel, das Konzept Topology hat die folgende Auftrittshäufigkeiten in den fünf Lernobjekte (LO_1 bis LO_5):

	LO_1	LO_2	LO_3	LO_4	LO_5
h	0	4	3	7	2

Das bedeutet, dass das Konzept Topology nicht im LO_1 vorkommt, aber 4-mal im LO_2, 3-mal im LO_3, 7-mal im LO_4 und 2-mal im LO_5. Zuerst wird die Position bestimmt, die das Lernobjekt für das jeweilige Konzept hat. Für eine gegebene Schranke der Position d, z.B.: $d = 1$, wird das Konzept Topology nur dem Lernobjekt LO_4 zugeordnet, da LO_4 die größte Auftrittshäufigkeit für dieses Konzept hat. Für $d = 2$ wird das Konzept den Objekten LO_4 und LO_2 zugeordnet, da die beiden LO die beiden größten Auftrittshäufigkeiten für dieses Konzept besitzen. h ist die Auftrittshäufigkeit des Konzeptes im Lernobjekt und d ist eine Schranke, die festlegt, in wieweit dieses Konzept einem entsprechenden Lernobjekt zugeordnet wird oder nicht.

Die Rollen werden immer alle ohne Ranking in die OWL-Datei übernommen.

Die semantische Annotation eines LO entspricht nun den relevanten Konzepten (nach dem Ranking) und den Rollen aus den jeweiligen Datenquellen:

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

m ist die Nummer des relevanten Konzeptes und d ist die Schranke für das durchgeführte Ranking. Das Ergebnis dieses Prozesses ist eine gültige DL-Beschreibung, ähnlich wie in Abbildung 2 dargestellt. Komplexe DL-Beschreibungen wie $\exists R.(A \sqcap \exists S.(B \sqcap A))$, ($A, B$ sind Konzepte, R, S sind Rollen) und Negationen $\neg A$ werden nicht erkannt und extrahiert. Nur einfache DL werden durch diesen Algorithmus erzeugt.

4 Versuchsvorbereitung

Die DL der Anfrage und die DL der Annotation werden einer semantischen Suchmaschine zugeführt, wie sie in der Veröffentlichung [KLM07] beschrieben ist. Die Suchmaschine berechnet die Ähnlichkeit eines Objektes der OWL-DL Annotation und der DL der Anfrage eines Suchenden. Die Maschine bestimmt die semantische Ähnlichkeit zwischen der Anfrage und der semantischen Beschreibung.

Der Spracherkenner wird in einem 15 Minuten dauernden Training auf den Sprecher eingestellt. Zusätzlich werden einige Domain-Wörter aus den Vortragsfolien in einem weiteren 15-minütigen Training in das Spracherkennungssystem aufgenommen. Die gesamte Trainingsphase des Spracherkenners beträgt also insgesamt 30 Minuten. Eine Wortgenauigkeit (*word-accuracy*) von ungefähr 60% des erzeugten Textstromes wird gemessen. Das Stemming in der Vorverarbeitung wird mit einem Porter-Stemmer durchgeführt [Por80].

Für unsere Versuche haben wir eine Vorlesung zum Thema “Internetworking” gewählt. Diese Vorlesung ist 100 Minuten lang und beinhaltet 62 Folien. Da jede Folie ein Multimedia Lernobjekt darstellt, stehen für die Versuche 62 Objekte zur Verfügung. Der Vortragende spricht über jede Folie ungefähr 1,5 Minuten. Diese Videosegmente stellen die LO dar.

Um die semantische Suche innerhalb dieser 62 Objekte zu testen, wurden von einer Expertengruppe 107 Fragen über das Thema Internetworking erstellt. Es handelte sich hierbei

um solche Fragen, wie sie Studenten an diese Vorlesung stellen könnten, z.B.: “Was ist eine IP-Adresse?”. Für jede Frage bestimmten die Experten genau ein passendes LO (Videosegment) als Goldstandard. Das bedeutet, dass es für jede Frage nur ein richtiges LO aus den 63 möglichen Objekten gibt.

Das Retrieval-Maß *recall* (R) beschrieben in [BYRN99] wird für die Evaluation der Ergebnisse herangezogen. Der Top-Recall Wert R_1 (R_5 oder R_{10}) wertet nur den ersten (oder fünften oder zehnten) Treffer des Ergebnisses aus. Der *reciprocal rank* (MRR), beschrieben in [Voo99], wird genutzt, um die Qualität der Ergebnisse besser einzuordnen. Ein MRR-Wert von 0,5 kann dahingehend interpretiert werden, dass im Durchschnitt der zweite Treffer aus der Liste die Frage beantwortet. Der MRR-Wert ist definiert:

$$MRR = \frac{1}{N} \sum_{i=1}^{N} (\frac{1}{r_i})$$

N ist die Anzahl der Fragen. r_i ist die Position der korrekten Antwort der Frage i in der Ergebnisliste. MRR_5 bedeutet, dass nur die fünf ersten Antworten mit in die Bewertung einfließen.

OWL-Dateien aus den Folien (F), den Transkripten des Spracherkenners (T), den korrigierten Transkripten (PT) und die Kombination von diesen, werden automatisch generiert. Zusätzlich wird eine manuell erstellte OWL-Datei (M) als Referenz hinzugezogen. Diese manuelle Annotation wurde von einem Expertenteam, bestehend aus drei Mitgliedern, entwickelt.

Die folgende Schreibweise wird für die unterschiedlichen Kombinationen verwendet:

$$[< source >]_d$$

$< source >$ steht für die Datenquelle (F, T, oder PT) und $< d >$ steht für die im vorangehenden Abschnitt beschriebene Schranke für die Auftrittshäufigkeiten h. Bei $d = 0$ werden **alle** Konzepte verwendet und den entsprechenden LO zugeordnet. Wird $d = 2$ gesetzt, wird das jeweilige Konzept nur den beiden LO zugeordnet, bei denen die Auftrittshäufigkeiten h des Konzeptes am größten sind. Z.B.: bedeutet $[T+F]_2$, dass die Konzepte aus dem fehlerbehafteten Transkript des Spracherkenners (T) und von den Folien (F) zuerst kombiniert werden (Vereinigungsmenge) und dann das Ergebnis verwendet wird, um das Ranking durchzuführen.

5 Durchgeführte Tests und Ergebnisse

Ausgehenden von diesen Annotationen wurden zwei Tests durchgeführt:
Der **erste Test** (Tabelle 1) analysiert, welche Datenquelle (F, T, PT) das beste Ergebnis durch die semantische Suchmaschine liefert. Hier zeigt sich erwartungsgemäß, dass das beste Ergebnis durch die manuell erstellte semantische Beschreibung (M) erreicht wird. Das Ergebnis lautet: 70% für R_1 und 82% für R_5. Betrachtet man nun die komplett automatisch generierten semantischen Beschreibungen (Datenquelle T und F), so erhält man

	R_1	R_2	R_3	R_4	R_5	R_{10}	MRR_5
Zeit	1,5 Min.	3 Min.	4,5 Min.	6 Min.	7,5 Min.	15 Min.	-
LO (Folien)	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	10 (10)	-
M	70	78	79	81	82	85	75
$[F]_0$	32	49	52	58	64	70	44
$[T]_2$	14	23	26	30	31	35	21
$[PT]_2$	25	33	37	40	41	44	31
$[T+F]_2$	36	42	46	50	52	64	42
$[PT+F]_2$	32	43	48	49	51	69	40

Tabelle 1: Maximale Zeiten, Recall-Werte und der MRR-Wert des **ersten Tests** (%).

für das System $[T]_2$ schlechtere Werte. Hier wird eine Frage in 14% der Fälle beim ersten Treffer korrekt beantwortet und in 31% der Fälle erhält der Lernende die richtige Antwort, wenn er sich die ersten fünf Treffer anschaut. Diese Werte können durch die Verwendung eines korrigierten Transkripts oder Verwendung eines besseren Spracherkenners leicht verbessert werden. In diesem Fall $[PT]_2$ wird ein MRR-Wert von 31% im Gegensatz zu einem MRR-Wert von 21% bei $[T]_2$ erreicht. Das Maximum der richtig beantworteten Fragen liegt bei einer Schranke von $d = 2$. Bei der Annotation mit Hilfe der Folien hat das beschriebene Ranking keinen messbaren Effekt. Festzustellen ist ebenso, dass die Folien die meisten Informationen enthalten und dass die Kombination von Folien mit Transkripten zu keinen besseren Resultaten führt.

Im **zweiten Test** (Tabelle 2) wird die zeitliche Abfolge der LO (eine Folie wird nach der anderen aufgelegt) berücksichtigt. Die Folien sind chronologisch in der Zeit verteilt und die benachbarten Folien beinhalten meistens thematisch verwandte LO. Die Antworten, die von der semantischen Suchmaschine geliefert werden, streuen um das richtige LO. Wird dieser Umstand berücksichtigt und eine Toleranz von einem LO vor und nach dem gesuchten LO akzeptiert, so steigt der MRR-Wert von $[T]_2$ um 15% und bei $[PT]_2$ sogar um 21% an. Die Hälfte der Fragen werden bereits mit den ersten drei LO richtig beantwortet. Dazu muss sich der Lernende 13,5 Minuten Video (3 Video-Objekte) anschauen. Auf Grund der Toleranz von plus/minus einem LO muss der Lernende ungefähr 4,5 Minuten Video pro LO ansehen (anstatt 1,5 Minuten); er befindet sich damit aber bereits innerhalb desselben Sinnzusammenhangs.

6 Bewertung und Einordnung

Mit Hilfe der völlig automatisch erstellten semantischen Beschreibung (T) (zweiter Test) wird die gestellte Frage in 22% ($[T]_2$) der Fälle mit dem ersten Treffer korrekt beantwortet. In 50% der Fälle werden die Fragen mit den ersten drei Treffern der Ergebnisliste korrekt beantwortet. Diese Werte können durch die Verwendung eines besseren Spracherkenners oder durch die nachträgliche Korrektur der Transkripte verbessert werden. Durch

	R_1	R_2	R_3	R_4	R_5	MRR_5
Zeit	4,5 Min.	9	13,5 Min.	18 Min.	22,5 Min.	-
LO(Folien)	1 (3)	2 (6)	3 (9)	4 (12)	5 (15)	-
$[F]_0$	42	57	62	66	70	53
$[T]_2$	22	43	50	55	56	36
$[PT]_2$	43	54	62	64	65	52
$[T+F]_2$	47	51	53	59	62	52
$[PT+F]_2$	43	54	65	66	70	53

Tabelle 2: Maximale Zeiten, Recall-Werte und der MRR-Wert des **zweiten Tests** (%).

diese Maßnahmen könnte theoretisch für die Kombination $[PT]_2$ 62% aller Fragen richtig beantwortet werden. Werden die Folien sowie ein perfektes Sprachtranskript verwendet ($[PT + F]_2$), so werden 65% der Fragen durch die drei ersten LO beantwortet. Ein LO hat jetzt inklusive der erlaubten Toleranz eine Länge von 4,5 Minuten (siehe Ergebnis zweiter Test). Das heißt, dass die Frage eines Lernenden, der im Schnitt $3 * 4,5 = 13,5$ Minuten Videosequenzen anschaut, zu 65% beantwortet wird.

Wie aber läßt sich dieses Ergebnis im Verhältnis zu denjenigen von bereits entwickelten, anderen Verfahren bewerten? Hier zeigt sich, dass die Vergleichbarkeit der Ergebnisse auf diesem Forschungsgebiet ein generelles Problem darstellt, da keine Referenzdatensätze (Videos, Transkripte, Fragen mit den passenden Antworten u.s.w.) existieren. Mit unserer Vorgehensweise am ehesten vergleichbar sind die Ergebnisse von [CFRN05]. Das hier vorgestellte System liefert einen MRR_5 Wert von 56%/62%. Der signifikante Unterschied ist jedoch, dass das beschriebene Verfahren schon vorsegmentierte Videos verwendet, während wir in unseren Versuchen diese Segmentation durch die Folienumbrüche automatisch erzeugen. Auch verwendet [CFRN05] nur 30 Fragen und die beschriebene Trainingsphase hat keinen unerheblichen Einfluss auf die Ergebnisse.

7 Fazit

In dieser Veröffentlichung wurde ein Algorithmus zur automatischen Erstellung semantischer Annotation von Videos vorgestellt. Der Algorithmus extrahiert Rollen und Konzepte aus den entsprechenden Textquellen. Die Annotationen werden anschließend in eine maschinenlesbare Form in einem OWL-Format gespeichert. Eine komplett automatische Erstellung von OWL-Dateien wurde präsentiert. Diese Methode kann die Arbeit eines Administrators verringern, der die Vorlesung annotieren muss. Es wurden zwei Tests mit jeweils vier Annotationen aus verschiedenen Quellen durchgeführt. Drei erzeugte Annotationen mit den Datenquellen Folien, Transkripte und korrigierte Transkripte wurden mit einer manuellen Annotation verglichen. Die Untersuchungen zeigten, dass knapp **zwei Drittel** aller Fragen durch eine semantische Suchmaschine, die die automatischen Annotationen mit Hilfe des vorgestellten Algorithmus verwendet, beantwortet werden können.

Das beschriebene Verfahren stellt damit eine kostengünstige und effektive Handhabung der automatischen semantischen Annotation vor, die insbesondere für Bildungseinrichtungen einen großen Mehrwert darstellen kann: Sie erleichtert sowohl Lehrenden als auch Lernenden die Arbeit mit digitalem Vorlesungsmaterial und optimiert die Möglichkeiten und v.a. die Praktikabilität des E-Learnings auf diese Weise deutlich.

Die erzeugten semantischen Annotationen sind einfache Beschreibungen der Lernobjekte. In unseren weiteren Forschungen wird untersucht, in wieweit komplexere Annotationen noch bessere Ergebnisse erzielen. Hierbei könnte der zeitliche Ablauf, die Einordnung in den Gesamtzusammenhang oder die Implementierung von Mustererkennungsverfahren zielführend sein. Ebenso wird untersucht, wie die Segmentatierung der Vorlesung in Lernobjekten unabhängig von den Folien automatisch bestimmt werden kann.

Die von uns verwendeten Daten können beim Autor angefordert und für weitere Forschungen auf diesem Gebiet genutzt werden.

Literatur

[BBT+06] Marco Bertini, Alberto Del Bimbo, Carlo Torniai, Rita Cucchiara und Costantino Grana. MOM: Multimedia Ontology Manager. A Framework for Automatic Annotation and Semantic Retrieval of Video Sequences. In *ACM SIGMM*, Seiten 787–788, 2006.

[BCM+03] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi und Peter F. Patel-Schneider, Hrsg. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[BYRN99] Ricardo A. Baeza-Yates und Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[CFRN05] Jinwei Cao, Jose Antonio Robles Flores, Dmitri Roussinov und Jay Nunamaker. Automated Question Answering From Lecture Videos: NLP vs. Pattern Matching. In *HICSS '05: Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 1*, Seite 43, Washington, DC, USA, 2005. IEEE Computer Society.

[EHLS06] Michael Engelhardt, Arne Hildebrand, Dagmar Lange und Thomas C. Schmidt. Reasoning about eLearning Multimedia Objects. In *International Workshop on Semantic Web Annotations for Multimedia (SWAMM)*, 2006.

[HK05] Alexander Haubold und John R. Kender. Augmented Segmentation and Visualization for Presentation Videos, 2005.

[HKW02] Wolfgang Hürst, Thorsten Kreuzer und Marc 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)*, Seiten 135–143, 2002.

[KLM07] Naouel Karam, Serge Linckels und Christoph Meinel. Semantic Composition of Lecture Subparts for a Personalized e-Learning. In *European Semantic Web Conference*, Jgg. 4519 of *Lecture Notes in Computer Science*, Seiten 716–728, 2007.

[LM06] Serge Linckels und Christoph Meinel. Resolving Ambiguities in the Semantic Interpretation of Natural Language Questions. In *Intelligent Data Engineering and Automated Learning (IDEAL)*, Jgg. 4224 of *LNCS*, Seiten 612–619, 2006.

[MSMV04] Robert Mertens, Holger Schneider, Olaf Mller und Oliver Vornberger. Hypermedia Navigation Concepts for Lecture Recordings. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, Seiten 2480–2847, 2004.

[NWP03] Chong-Wah Ngo, Feng Wang und Ting-Chuen Pong. Structuring Lecture Videos for Distance Learning Applications. In *Multimedia Software Engineering*, Seiten 215–222, 2003.

[Por80] M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[RLM07] Stephan Repp, Serge Linckels und Christoph 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*, Seiten 19–26. ACM, 2007.

[RM06a] Stephan Repp und Christoph Meinel. Segmenting of Recorded Lecture Videos - The Algorithm VoiceSeg. In *Proceedings of the 1th Signal Processing and Multimedia Applications*, Seiten 317–322, August 2006.

[RM06b] Stephan Repp und Christoph Meinel. Semantic Indexing for Recorded Educational Lecture Videos. In *4th IEEE Conference on Pervasive Computing and Communications Workshops (PerCom 2006 Workshops), 13-17 March 2006, Pisa, Italy*, Seiten 240–245. IEEE Computer Society, 2006.

[RWSM07] Stephan Repp, Jörg Waitelonis, Harald Sack und Christoph Meinel. Segmentation and Annotation of Audiovisual Recordings Based on Automated Speech Recognition. In *Intelligent Data Engineering and Automated Learning - IDEAL 2007, 8th International Conference, Birmingham, UK, December 16-19*, Jgg. 4881 of *Lecture Notes in Computer Science*, Seiten 620–629. Springer, 2007.

[Sch93] Renate A. Schmidt. Terminological Representation, Natural Language & Relation Algebra. In *German AI Conference (GWAI)*, Jgg. 671 of *LNCS*, Seiten 357–371, 1993.

[SW06] Harald Sack und Jörg Waitelonis. Integrating Social Tagging and Document Annotation for Content-Based Search in Multimedia Data. In *Semantic Authoring and Annotation Workshop (SAAW)*, 2006.

[Voo99] Ellen M. Voorhees. The TREC-8 Question Answering Track Report. In *TREC*, 1999.

[W3C04] World Wide Web Consortium W3C. *OWL Web Ontology Language*. `http://www.w3.org/TR/owl-features/`, 2004.

[WPS+04] Patrick Wolf, Wolfgang Putz, Avare Stewart, Arnd Steinmetz, Matthias Hemmje und Erich Neuhold. LectureLounge – experience education beyond the borders of the classroom. *International Journal on Digital Libraries*, 4(1):39–41, 2004.

[YOA03] Natsuo Yamamoto, Jun Ogata und Yasuo Ariki. Topic Segmentation and Retrieval System for Lecture Videos Based on Spontaneous Speech Recognition. In *European Conference on Speech Communication and Technology*, Seiten 961–964, 2003.