

Audiosuche: Information Retrieval in multimedialen Datenbanken

W. Macherey, H. Ney

Lehrstuhl für Informatik VI, RWTH Aachen
Ahornstraße 55, 52056 Aachen, Germany
wmach@informatik.rwth-aachen.de

1 Zielsetzung

Im Rahmen des Projektes „Multimedia NRW: Die virtuelle Wissensfabrik“ wurde von Seiten des Lehrstuhls für Informatik VI der RWTH Aachen eine erste Version eines Systems zur Dokumentensuche in multimedialen Datenbanken entwickelt und evaluiert. Dabei stand die automatische Erkennung und Verarbeitung von Audiodaten sowie die Entwicklung geeigneter Retrieval-Methoden, die sich robust gegenüber Fehlern in der Erkennung verhalten, im Vordergrund.

Während des Projektes wurde hierzu am Lehrstuhl für Informatik VI zunächst ein textbasiertes Information-Retrieval (IR)-System entwickelt, mit dem sich Anfragen in natürlich geschriebener Sprache an eine Datenbank richten lassen. Das IR-System ermittelt zu jeder Anfrage eine Menge von Dokumenten, die anschließend dem Benutzer in einer Rangfolge gemäß ihrer angenommenen Relevanz präsentiert werden.

In einer zweiten Phase wurde dieses System durch Integration des Aachener Spracherkenners für großen Wortschatz im Hinblick auf die Audiosuche erweitert. Der Erkenner transkribiert hierbei automatisch die Sprachdokumente und stellt sie der Datenbank für die weitere Verarbeitung zur Verfügung.

Das entwickelte IR-System wurde an der Aufgabe des *Spoken-Document-Retrieval* (SDR) *Tracks* der *Text Retrieval Conference 7* (TREC-7) 1998 auf dem Hub-4 Korpus getestet.

2 Einführung

Angesichts der steigenden Informationsflut an unstrukturierten Daten in multimedialen Datenbanken und auch im Internet gewinnen die Methoden des Information-Retrievals zunehmend an Bedeutung. Bereits heute gibt es sog. „Search Engines“, die die im Internet zugänglichen Text-Dokumente nach verwertbaren Informationen absuchen und entsprechende Verweise anlegen. Um die darin enthaltenen Informationen dem Benutzer zugänglich zu machen, werden effiziente Verfahren benötigt, um die Dokumente und die in ihnen enthaltene Information zu strukturieren, zu klassifizieren und zu filtern. Ein wichtiger Gesichtspunkt hierbei ist, daß die Information häufig nicht mehr als geschriebener Text, sondern nur als Audiosignal vorliegt, z.B. für Nachrichtensendungen, Interviews, Diskussionsrunden usw. Hier ergibt sich die Notwendigkeit, aus diesen multimedialen Datenbanken, d.h. insbesondere Sprach- und Videoaufzeichnungen, den gesprochenen Text zu erkennen und zu extrahieren und so erst ein effizientes Information-Retrieval zu ermöglichen.

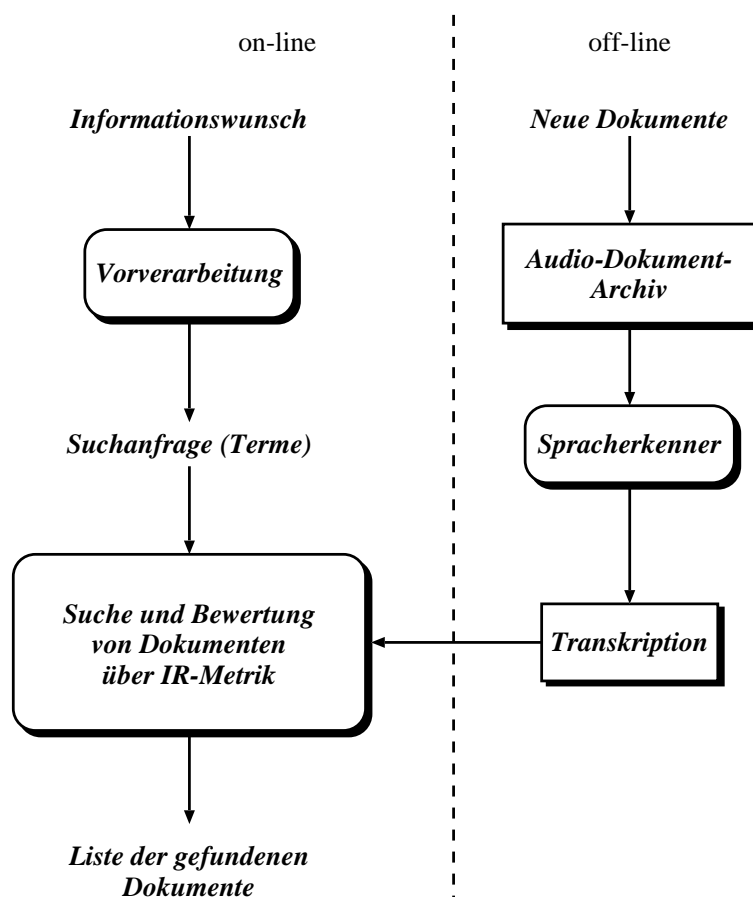


Abbildung 1: Architektur eines Information-Retrieval-Systems für die Suche in Audio-Dokumenten.

Im Rahmen dieses Projektes wurde von Seiten der RWTH eine Spracherkennungskomponente zur automatischen Transkription von Nachrichtensendungen sowie ein Information-Retrieval-System für die Volltextsuche in den durch die Erkennung aufbereiteten Tondokumenten entwickelt. Die automatische Transkription der Audio-Dokumente erfolgt dabei aufgrund der hohen Rechenzeitanforderungen im *offline*-Betrieb, das heißt die Tondokumente werden vorab durch das Spracherkennungssystem transkribiert und die so generierten Transkriptionen für die spätere Volltextsuche in einer Datenbank abgelegt. Die Volltextsuche erfolgt unter Verwendung von Methoden des Information-Retrievals in einem von der Transkriptionsaufgabe unabhängigen Prozeß. Abbildung 1 zeigt schematisch den Aufbau eines Information-Retrieval-Systems für die Suche in Audio-Dokumenten. Im folgenden werden die beiden Komponenten näher beschrieben.

3 Das Hub-4 Erkennungssystem

Zur automatischen Transkription der Audio-Dokumente wurde der Aacheener Spracherkennung eingesetzt, der für die sprecherunabhängige Erkennung kontinuierlich gesprochener Sprache bei großem Wortschatz ausgelegt ist [Ney et al. 98a]. Der Erkennungsbasiert auf einem zeitsynchronen Suchalgorithmus unter Verwendung wortabhängiger Baumkopien wobei die Worthypothesen mittels eines integrierten Trigramm-Sprachmodells rekombiniert werden. Um den Suchaufwand zu reduzieren, wird neben den üblichen Pruningtechniken, zu denen das akustische Pruning und das Histogramm-Pruning gehören, zusätzlich ein Sprachmodell-look-ahead Pruning verwendet [Ortmanns et al. 98]. Die Segmentierung von Sprach- und Nicht-Sprachanteilen im akustischen Signal erfolgt auf Grundlage eines eigens hierzu trainierten Gaußschen Mischverteilungsmodells. Training und Erkennung verwenden cepstrale Merkmalsvektoren, von denen jeder aus 16 MFC-Koeffizienten zusammengesetzt ist. Von den resultierenden Merkmalsvektoren werden jeweils neun aufeinanderfolgende Vektoren unter Verwendung einer linearen Diskriminanzanalyse (LDA) auf einen 45-komponentigen Merkmalsvektor abgebildet. Das System arbeitet unter Verwendung Gaußscher Mischverteilungsdichten und einem über alle Zustände gepoolten Varianzvektor.

4 Die IR-Komponente

Die entwickelte IR-Systemkomponente ist für die Verarbeitung und Suche in englischsprachigen Texten ausgelegt. Die Dokumente können vorverarbeitet und automatisch indiziert werden. Hierzu werden zunächst sämtliche Funktionswörter aus den Texten entfernt, die für eine Dokumentensuche keine Bedeutung besitzen (*stopping*). Die verwendete Stoppliste umfaßt 319 der häufigsten englischen Funktionswörter. Als Indexterme werden die Wortstämme der verbleibenden Wörter verwendet, wobei zur Wortstambildung der Algorithmus nach Porter [Porter 80] eingesetzt wird.

Im folgenden bezeichne d ein Dokument, q eine Anfrage und t einen Indexterm. Ein Dokument d ist eine Folge von Indextermen der Form $d = [t_1^d, \dots, t_{N(d)}^d]$. Analog sei für eine Anfrage $q = [t_1^q, \dots, t_{N(q)}^q]$. Jedem Indexterm t eines Dokumentes d wird nun ein Gewicht $\tilde{w}(t, d)$ zugewiesen, das gemäß [Choi et al. 98] als Verhältnis aus dem Logarithmus der absoluten Termfrequenz (*term frequency*) $\text{tf}(t, d)$ und der mittleren Termfrequenz $\overline{\text{tf}}(d)$ definiert ist:

$$\tilde{w}(t, d) = \begin{cases} \left[1 + \log(\text{tf}(t, d))\right] / \left[1 + \log(\overline{\text{tf}}(d))\right] & \text{falls } t \in d \\ 0 & \text{falls } t \notin d \end{cases} \quad (1)$$

Die absolute Termfrequenz $\text{tf}(t, d)$ gibt an, mit welcher Häufigkeit der Term t im Dokument d auftritt:

$$\text{tf}(t, d) = \sum_{n=1}^{N(d)} \delta(t, t_n^d) \quad (2)$$

Die mittlere Termfrequenz $\overline{\text{tf}}(d)$ ist definiert als die durchschnittliche Zahl der Vorkommen von Termen t in einem Dokument d :

$$\overline{\text{tf}}(d) = \frac{\sum_t \text{tf}(t, d)}{\sum_{t: \text{tf}(t, d) > 0} 1} \quad (3)$$

Die Logarithmen in Gleichung (1) verhindern nun, daß Dokumente mit sehr kleinen Termfrequenzen durch Dokumente mit großen Termfrequenzen überdeckt werden. Zur Bestimmung des endgültigen Gewichts $w(t, d)$ wird der Ausdruck in Gleichung (1) durch eine Linearkombination aus einem Pivot-Wert k und der Zahl der Singletons $n_1(d)$ für ein gegebenes Dokument d dividiert:

$$w(t, d) = \frac{\tilde{w}(t, d)}{0.8 \cdot k + 0.2 \cdot n_1(d)} \quad \text{mit } n_1(d) := \sum_{t: \text{tf}(t, d) = 1} 1 \quad (4)$$

Für die Wahl des Pivot-Wertes wird

$$k = \frac{1}{D} \sum_d \sum_{t: \text{tf}(t, d) = 1} 1 \quad (5)$$

gewählt, wobei D die Gesamtzahl aller Dokumente in der Datenbank bezeichnet. In ähnlicher Weise werden nun die Termgewichte $w(t, q)$ bzgl. einer Anfrage q bestimmt. Diese Gewichte sind nach Gleichung (6) definiert als das Produkt aus dem Logarithmus der Termfrequenz $\text{tf}(t, q)$ und der inversen Dokumenthäufigkeit (*inverse document frequency*) $\text{idf}(t)$:

$$w(t, q) = \left(1 + \log(\text{tf}(t, q))\right) \cdot \text{idf}(t) \quad (6)$$

Hierbei ist $\text{idf}(t)$ durch den folgenden Ausdruck bestimmt:

$$\text{idf}(t) = \log \frac{D}{\text{df}(t)} \quad \text{mit } \text{df}(t) := \sum_d \sum_{n=1}^{N(d)} \delta(t, t_n^d) \quad (7)$$

Anfragen an das System werden mit einer Liste der für relevant befundenen Dokumente beantwortet, wobei sich die Relevanzbeurteilung (*retrieval status value*, RSV) eines Dokumentes d bzgl. einer Anfrage q als „Abstand“ zwischen den Gewichtsvektoren $\mathbf{w}(d) \equiv (w(t_1^d, d), \dots, w(t_{N(d)}^d, d))^\top$ und $\mathbf{w}(q) \equiv (w(t_1^q, q), \dots, w(t_{N(q)}^q, q))^\top$ ergibt:

$$\text{RSV}(q, d) = \sum_{n=1}^{N(q)} w(t_n^q, q) \cdot w(t_n^q, d) \quad (8)$$

5 Bewertung der Retrieval-Effektivität

Um die Retrieval-Effektivität ermitteln zu können, werden sog. *Recall-Precision-Graphen* bestimmt, die den Anteil der relevanten Dokumente unter allen ermittelten Dokumenten (*precision*) in Beziehung setzen zu dem Verhältnis aus gefundenen relevanten Dokumenten und insgesamt in der Datenbank vorhandenen relevanten Dokumenten (*recall*) [Schäuble 97].

5.1 Berechnung von *Recall-Precision-Graphen*

Im folgenden bezeichne \mathcal{D} eine Menge von Dokumenten. \mathcal{Q} sei eine Menge von Anfragen. Für jede Anfrage $q \in \mathcal{Q}$ zerfällt die Menge \mathcal{D} in zwei Teilmengen $\mathcal{D}^{\text{rel}}(q)$ und $\mathcal{D}^{\text{irr}}(q)$. Hierbei bezeichnet $\mathcal{D}^{\text{rel}}(q)$ die Menge der bzgl. q relevanten Dokumente und $\mathcal{D}^{\text{irr}}(q)$ die Menge der bzgl. q irrelevanten Dokumente.

Die Dokumente $d \in \mathcal{D}$ werden nun für jede Anfrage $q \in \mathcal{Q}$ gemäß der Bewertungsfunktion $\text{RSV}(q)$ in absteigender Reihenfolge angeordnet. Unter der vereinfachenden Annahme, daß sämtliche Retrieval-Status-Werte verschieden sind, existiert genau eine Permutation π , die die Menge $\mathcal{D} = \{d_1, \dots, d_D\}$ bijektiv auf sich selbst abbildet

$$\pi : \begin{cases} \mathcal{D} & \rightarrow \mathcal{D} \\ d & \mapsto \pi(d) \end{cases}, \quad (9)$$

so daß gilt:

$$\text{RSV}(q, d_{\pi(1)}) > \text{RSV}(q, d_{\pi(2)}) > \dots > \text{RSV}(q, d_{\pi(D)}) \quad (10)$$

Für die Definition der *Recall*- und *Precision*-Werte wird ferner noch der Begriff der *Antwortmenge* benötigt. Die Antwortmenge $\mathcal{D}_r(q)$ ist definiert als die Menge der Dokumente mit den r größten Retrieval-Status-Werten bzgl. q :

$$\mathcal{D}_r(q) := \{d_{\pi(1)}, \dots, d_{\pi(r)}\} \quad (11)$$

Bezüglich einer Anfrage q sind die *Recall*- und *Precision*-Werte dann wie folgt definiert:

$$\begin{aligned} \text{Recall-Wert:} & \quad \phi_r(q) := \frac{|\mathcal{D}_r^{\text{rel}}(q)|}{|\mathcal{D}^{\text{rel}}(q)|} \\ \text{Precision-Wert:} & \quad \psi_r(q) := \frac{|\mathcal{D}_r^{\text{rel}}(q)|}{|\mathcal{D}_r(q)|} \end{aligned} \quad (12)$$

Hierbei gilt $\mathcal{D}_r^{\text{rel}}(q) := \mathcal{D}^{\text{rel}}(q) \cap \mathcal{D}_r(q)$, d.h. die Menge $\mathcal{D}_r^{\text{rel}}(q)$ identifiziert unter den r gefundenen Dokumenten den für die Anfrage relevanten Anteil.

Um nun jedem *Recall*-Wert $\phi \in [0, 1]$ einen *Precision*-Wert zuzuordnen, wird die folgende Funktion definiert:

$$\Psi_q(\phi) := \max\{\psi_r(q) \mid \phi_r(q) \geq \phi\} \quad (13)$$

Bestimmt man nun das arithmetische Mittel der Funktion $\Psi_q(\phi)$ für jedes $\phi \in [0, 1]$, so ergibt sich der sog. *Recall-Precision-Graph* als Graph der Funktion

$$\Psi(\phi) := \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \Psi_q(\phi). \quad (14)$$

6 Experimente und Diskussion der Ergebnisse

Das entwickelte Information-Retrieval-System wurde auf dem SDR-Track der TREC-7 Evaluierung für den Hub-4 Korpus (98er Trainings-Daten) getestet [Robinson et al. 99]. Die Datenbasis umfaßt über 100 Stunden Tonmaterial amerikanischer Rundfunk- und Fernsehnachrichten, die in 2866 Einzeldokumente unterschiedlicher Länge zerfallen. Für die Lernphase wurden ca. 43h der insgesamt 76h des Trainingsmaterials der 96er und 97er Trainingsdaten des Hub-4 Korpus verwendet.

Zur Evaluierung wurden 23 Testanfragen an das System gestellt, für die jeweils die Menge der relevanten Dokumente zu ermitteln war. Die Testläufe wurden sowohl auf den manuell transkribierten Daten (Originaltranskriptionen) als auch auf den automatisch erzeugten Transkriptionen durchgeführt. Als Maß für die Retrieval-Effektivität wurden die oben eingeführten *Recall-Precision-Graphen* verwendet. Die auf den 98er-Trainingsdaten erzielte Erkennungsfehlerrate sowie die Effektivität der Retrieval-Ergebnisse sind aus Tabelle 1 bzw. aus Abbildung 2 ersichtlich.

Tabelle 1: Wortfehlerrate (*word error rate*) auf den 98er Trainingsdaten des Hub-4 Korpus und mittlere Präzisionsrate (*mean average precision*) auf den Originaltranskriptionen (*text*) sowie auf den automatisch generierten Transkriptionen (*speech*).

| Corpus | | Hub-4 Train. |
|---------------------------|--------|--------------|
| Duration[h] | | 107 |
| Word Error Rate[%] | | 32.5 |
| Mean Average Precision[%] | text | 46.56 |
| | speech | 42.01 |

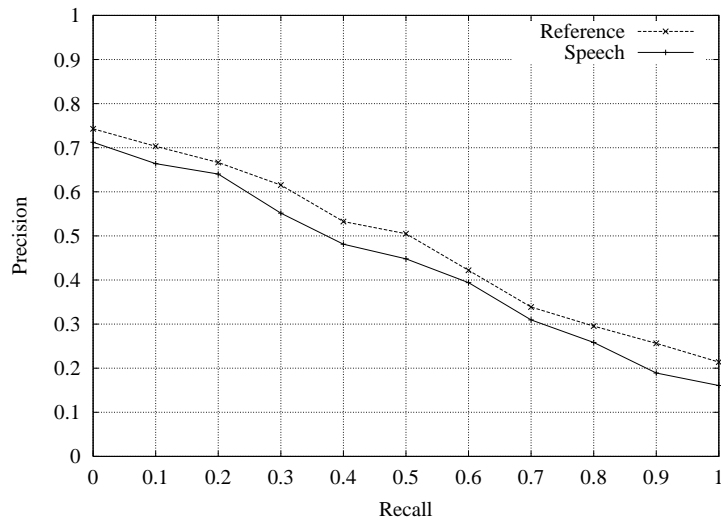


Abbildung 2: Recall-Precision-Graph.

Die nachfolgende Tabelle zeigt exemplarisch einige typische Anfragen aus der TREC-7 Evaluierung sowie einen Auszug der hierzu gefundenen Dokumente (durchgeführt auf den Originaltranskriptionen). Das Symbol ✓ bedeutet dabei, daß es sich um ein relevantes Dokument handelt, während ✕ ein irrelevantes Dokument kennzeichnet.

Tabelle 2: Beispiele für Text-Retrieval-Anfragen auf den Originaltranskriptionen.

| | |
|---------------|---------------------------------------------------------------------------------------------------------------------------------------|
| Query: | <i>Do genes and/or hormones pay a role in causing or preventing cancer? If so, what is it?</i> |
| Doc.: | Also on health, a gene inherited by people with a rare intestinal disease could help scientists unlock some clues about cancer. ... ✓ |
| Query: | <i>What areas of the United States suffered from flooding?</i> |
| Doc.: | Checking our top stories, flash flood warnings are posted for parts of Georgia and the Carolinas Remnants of hurricane Danny, ... ✓ |
| Doc.: | The Australian government has declared a state of emergency in flood ravaged areas of North Queensland. ... ✕ |
| Query: | <i>What data is available on volcanic activity on the island of Montserrat?</i> |
| Doc.: | Many residents of Montserrat attended mass Sunday despite a deadly volcano blowing nearby. ... ✓ |

Tabelle 3: Beispiel für eine *Spoken-Document-Retrieval*-Anfrage auf den automatisch generierten Transkriptionen.

| | |
|----------------|-----------------------------------------------------------------------------------------------------------------------------|
| Query: | <i>What data is available on volcanic activity on the island of Montserrat?</i> |
| Spoken: | ... A volcano threatens a tiny Caribbean island, but only a few residents are fleeing. We'll tell you why when we return. ✓ |
| Recog.: | ... of volcanoes friends a tiny Caribbean island but only a few residents are fleeing. We'll tell you why when we return. ✓ |
| Spoken: | Many residents of Montserrat attended mass Sunday despite a deadly volcano blowing nearby. ... ✓ |
| Recog.: | Any residents of months the rot attended mass Sunday despite and deadly volcano blowing nearby. ... ✓ |

Tabelle 3 vermittelt einen Eindruck über die Suche in den automatisch transkribierten Dokumenten. Aufgeführt ist die zugrundeliegende Anfrage sowie Ausschnitte aus den Original-Dokumenten (Spoken) bzw. den erkannten Dokumenten (Recognized).

Ein Vergleich der mittleren Präzisionsrate (*mean average precision*) für ein Retrieval auf den Originaltranskriptionen mit den automatisch generierten Transkriptionen (siehe Tabelle 1) zeigt, daß sich die verwendete Retrieval-Metrik verhältnismäßig robust gegenüber Fehlern in der Erkennung verhält. So ergibt sich trotz einer Wortfehlerrate (*word error rate*) von 32.5% lediglich eine Relativverschlechterung von 9.8% bzgl. der mittleren Präzisionsrate, wenn man von der Volltextsuche auf den Originaltranskriptionen zur Textsuche auf den automatisch generierten Transkriptionen wechselt.

Tabelle 4 enthält exemplarisch für die Volltextsuche auf den Originaltranskriptionen neben den für relevant befundenen Dokumenten zusätzlich die berechneten Retrieval-Status-Werte. Zum Vergleich sind in Tabelle 5 die Retrieval-Status-Werte der gefundenen Dokumente auf den automatisch transkribierten Dateien aufgeführt.

Üblicherweise ist das für die Lernphase des Spracherkenners zur Verfügung stehende Trainingsmaterial nicht umfangreich genug, um damit die verwendeten statistischen Modelle sicher schätzen zu können. Die Konsequenz hieraus ist ein Verlust in der Performanz des Retrieval-Systems aufgrund von Fehlern in der Erkennung. Sieht man nun Methoden vor, die auch im laufenden Betrieb des Systems eine Fortführung des Trainings ermöglichen, so läßt sich damit die Performanz des Spracherkenners und somit auch die Effektivität der Retrieval-Komponente verbessern. Ein solches Verfahren ermöglicht das unüberwachte Lernen (sog. *unsupervised training*).

Tabelle 4: Beispiel für eine *Text-Retrieval*-Anfrage auf den Originaltranskriptionen des 98er Hub-4 Trainingskorpus. Die zweite Spalte enthält die Bewertungen der gefundenen Dokumente.

| | | |
|---------------|-----------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Query: | <i>What states have reached a settlement of suits vs tobacco companies?</i> | |
| Doc. | 14.62 | Florida's joined Mississippi as the second state to settle a court case against the big tobacco companies, ... ✓ |
| Doc. | 13.14 | The national multi-billion dollar tobacco settlement reached last month may not be a done deal, but the state of Mississippi is made sure it gets it's share of the money no matter what happens. ... ✓ |
| Doc. | 12.26 | The tobacco industry has agreed to pay the state of Mississippi 3.6 billion dollars to cover the cost of treating tobacco related illness. ... ✓ |
| Doc. | 12.12 | A couple of minutes ago, we reported that the big tobacco companies were in Congress today, pushing hard for a national settlement of all the lawsuits against them. ... ✗ |
| Doc. | 11.22 | In Washington, a California biotechnology company has pleaded guilty to illegally conspiring with a major tobacco company to grow tobacco with extra nicotine. ... ✗ |

Tabelle 5: Beispiel für eine *Spoken-Document-Retrieval*-Anfrage auf den automatisch generierten Transkriptionen des 98er Hub-4 Trainingskorpus. Die zweite Spalte enthält die Bewertungen der gefundenen Dokumente.

| | | |
|---------------|-----------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Query: | <i>What states have reached a settlement of suits vs tobacco companies?</i> | |
| Doc. | 12.97 | The tobacco industry has agreed to pay the state of Mississippi 3.6 billion dollars to cover the cost of treating tobacco related illness. ... ✓ |
| Doc. | 12.45 | The national multi-billion dollar tobacco settlement reached last month may not be a done deal, but the state of Mississippi has made sure gets its share of the money the matter what happens. ... ✓ |
| Doc. | 12.27 | Let's go , we reported the big tobacco companies were in Congress today, pushing hard for a national settlement of all lawsuits against them. ... ✗ |
| Doc. | 11.09 | In Washington, a California biotechnology company has pleaded guilty to illegally conspiring with the major tobacco company to grow tobacco with extra nicotine. ... ✗ |
| Doc. | 11.02 | Forester in Mississippi as a second state to settle a court case against the big tobacco companies, ... ✓ |

Von Seiten der RWTH wurden hierzu bereits erste Experimente auf den 98er Trainingsdaten des Hub-4 Korpus durchgeführt. Dabei bildete die automatisch generierte Transkription der 98er Daten die Grundlage für ein neu aufgesetztes Training. Mit den so trainierten Referenzen konnte auf dem Eval96 Korpus (Hub-4) eine Fehlerrate von 36.2% erzielt werden. Im Vergleich dazu betrug die Fehlerrate der mit den Originaltranskriptionen des 98er Korpus trainierten Modelle 34.9%. Der verhältnismäßig geringe Performanzverlust kann als Indiz für die Durchführbarkeit eines unüberwachten Lernens zur Verbesserung der Erkennungskomponente im laufenden Betrieb eines Retrieval-Systems betrachtet werden.

7 Zusammenfassung und Ausblick

Im Rahmen des Projektes „Multimedia NRW: Die virtuelle Wissensfabrik“ wurde von Seiten der RWTH Aachen eine Spracherkennungskomponente zur automatischen Transkription von Audio-Dokumenten sowie ein Information-Retrieval-System für die Volltextsuche in den durch die Erkennung aufbereiteten Tondokumenten entwickelt. Experimente auf den 98er Daten des Hub-4 Korpus zeigten auf dem SDR-Track der TREC-7 Evaluierung eine Retrieval-Effektivität von 46.56% auf den Originaltranskriptionen sowie 42.01% auf den automatisch transkribierten Dokumenten. Um die Performanz des Erkenners zu verbessern, wurden erste Experimente zu einem unüberwachten Lernen durchgeführt. Hierbei zeigte sich, daß die Erkennungsergebnisse für Parametersätze, die auf automatisch generierten Transkriptionen trainiert werden, bereits eine erste Basis für eine Fortsetzung des Trainings im laufenden Betrieb des Systems bilden. Für die Zukunft stehen die folgenden Arbeiten an:

- **Kompensation von Erkennungsfehlern**

Um Erkennungsfehler besser zu berücksichtigen, können zum einen Tabellen mit Phonem-Verwechslungswahrscheinlichkeiten eingesetzt und in die IR-Metrik einbezogen werden. Zum anderen kann man auch sogenannte Wortgraphen (statt der besten erkannten Wortfolge) verwenden. Durch einen Wortgraphen kann die Menge der wahrscheinlichsten Wortfolge-Alternativen kompakt dargestellt werden und als weitere Wissensquelle in die Retrieval-Metrik einfließen.

- **Entwicklung eines IR-Systems für die deutsche Sprache**

Für die Suche in Audio-Dokumenten bietet sich ferner die Entwicklung eines IR-Systems für die deutsche Sprache an. Da im Unterschied zum Englischen die deutsche Sprache die Bildung zahlreicher Komposita zuläßt, ergeben sich hier wichtige Aufgabenfelder, die die Vorverarbeitung der Dokumente betreffen. Insbesondere sind Algorithmen zu entwickeln, die ähnlich dem Porter-Algorithmus eine Reduktion der Wörter auf ihren Wortstamm ermöglichen sowie Verfahren, die die Komposita in die zugrundeliegenden Einzelwörter zerlegen.

- **Schnelles Training**

Aus Gründen der Handhabbarkeit und um den personellen Aufwand beim Aufbau einer Erkennungskomponente für ein IR-System möglichst gering zu halten, ist es erforderlich, Systeme zu entwickeln, die auch mit sehr begrenztem Trainingsmaterial in Betrieb gehen können. Um die Performanz derartiger Systeme zu verbessern, müssen Verfahren entwickelt werden, die eine Fortführung des Trainings im laufenden Betrieb ermöglichen. Die hierzu zählende Methode des unüberwachten Lernens kann durch den Einsatz von Konfidenzmaßen verbessert werden. Die Konfidenzmaße detektieren hierbei die korrekt erkannten Wörter aus einer automatisch generierten, aber möglicherweise fehlerbehafteten Transkription. Für das weiterführende Training werden dann nur diejenigen Segmente aus dem akustischen Signal verwendet, für die die Erkennung hinreichend konfident war. Auf diese Weise ist es möglich, die Modelle des Systems auch im laufenden Betrieb zu trainieren, was zu einer kontinuierlichen Verbesserung der Performanz führt.

8 Danksagung

Diese Arbeit wurde teilweise durch das Ministerium für Wissenschaft und Forschung (MWF) des Landes Nordrhein-Westfalen (NRW) im Rahmen des Forschungsverbundes „Multimedia NRW: Die virtuelle Wissensfabrik“ unter dem Kennzeichen IV A 3 - 107 028 97 gefördert.

Literatur

- [Beyerlein et al. 99] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, M. Pitz und A. Sixtus, *The Philips/RWTH System for Transcription of Broadcast News*, Proc. of DARPA Broadcast News Workshop (Herndon, VA), February 28-March 3 1999, S. 151–155.
- [Choi et al. 98] J. Choi, D. Hindle, J. Hirschberg, I. Magrin-Changnonleau, C. Nakatani, F. Pereira, A. Singhal und S. Whittaker, *An Overview of the AT&T Spoken Document Retrieval*, Broadcast News Transcription and Understanding Workshop (DARPA) (Lansdowne, VA), February 1998, S. 182–188.
- [Ney et al. 98a] H. Ney, L. Welling, S. Ortmanms, K. Beulen und F. Wessel, *The RWTH Large Vocabulary Speech Recognition System*, IEEE Int. Conf. on Acoustics, Speech and Signal Processing (Seattle, WA), May 1998, S. 853–856.
- [Ney et al. 98b] H. Ney, L. Welling, S. Ortmanms, K. Beulen und F. Wessel, *The RWTH Speech Recognition System and Spoken Document Retrieval*, 24th Annual Conference of the IEEE Industrial Electronics Society (IECON) (Aachen), September 1998, S. 2022–2027.
- [Ortmanms et al. 98] S. Ortmanms, A. Eiden und H. Ney, *Improved Lexical Tree Search for Large Vocabulary Speech Recognition*, IEEE Int. Conf. on Acoustics, Speech and Signal Processing (Seattle, WA), May 1998, S. 817–820.
- [Porter 80] M. F. Porter, *An Algorithm for Suffix Stripping*, July 1980, Programm 14(3), S. 130–137.
- [Robinson et al. 99] P. Robinson, E. Brown, J. Burger, N. Chinchor, A. Douthat, L. Ferro und L. Hirschman, *Overview: Information Extraction from Broadcast News*, DARPA Broadcast News Workshop (Herndon, VA), February 1999, S. 27–30.
- [Schäuble 97] Peter Schäuble, *Multimedia Information Retrieval*, Kluwer Academic Publishers, Norwell, MA, 1997.