In: Bekavac, Bernard; Herget, Josef; Rittberger, Marc (Hg.): Informationen zwischen Kultur und Marktwirtschaft. Proceedings des 9. InternationalenSymposiums fur Informationswissenschaft (ISI 2004), Chur, 6.-8.Oktober2004. Konstanz: UVK Verlagsgesellschaft mbH, 2004. S. 269 – 281

ASEMOS. Weiterentwicklung der Behandlung semantischer Heterogenität

Robert Strötgen

Universität Hildesheim
Institut für Angewandte Sprachwissenschaft
Marienburger Platz 22
D-31141 Hildesheim
stroetgen@uni-hildesheim.de

Zusammenfassung

Im Projekt *CARMEN* wurde unter anderem versucht, die Erweiterung von Anfragen an bibliografische Datenbanken hin zu Internet-Dokumenten zu verbessern. Dazu wurden verschiedene Ansätze zur Behandlung semantischer Heterogenität gewählt (Extraktion von Metadaten, Übersetzung von Anfragen unter Verwendung statistischer Relationen und intellektuell erstellter Cross-Konkordanzen). Im Nachfolgeprojekt ASEMOS weren einige offene Punkte weiter verfolgt. Dabei werden unter anderem Methoden des maschinellen Lernens eingesetzt, um die Qualität der Ergebnisse beim Wechseln zwischen Ontologien zu verbessern. In diesem Artikel werden die wesentlichen Konzepte und Umsetzungen dieses Ansatzes kurz dargestellt und Experimente zur Verbesserung beschrieben.

Abstract

The project *CARMEN* aimed among other goals at improving the expansion of searches in bibliographic databases into Internet searches. For this purpose a set of different approaches to the treatment of semantic heterogeneity (meta-data extraction, query translation using statistic relations and cross-concordances) were pursued. Subsequent to this project some open issues are proceeded in the follow-up project ASEMOS using machine learning technology to improve the results of Ontology Switching. This paper describes the main concepts and implementation of this approach and outlines the experiments to improve the approach.

Dieses Dokument wird unter folgender creative commons Lizenz veröffentlicht:
http://creativecommons.org/licenses/by-nc-sa/2.0/de/

1 Behandlung semantischer Heterogenität in *CARMEN*

In digitalen Bibliotheken als integrierten Zugängen zu in der Regel mehreren verschiedenen Dokumentsammlungen tritt Heterogenität in vielerlei Spielarten auf:

- Als technische Heterogenität durch das Zusammenspiel verschiedener Betriebs-, Datenbank- oder Softwaresysteme,
- als strukturelle Heterogenität durch das Auftreten verschiedener Dokumentstrukturen und Metadaten-Standards und schließlich
- als semantische Heterogenität, wenn Dokumente mit Hilfe unterschiedlicher Ontologien (hier verwendet im weiteren Sinn von Dokumentationssprachen wie Thesauri und Klassifikationen) erschlossen wurden oder aber Dokumente überhaupt nicht mit Metadaten ausgezeichnet wurden.

Semantische Heterogenität lässt sich behandeln, indem die Standardisierung von Metadaten (z.B. von der Dublin Core Metadata Initiative[1] oder das *Resource Description Framework*[2] (RDF) im Kontext des *Semantic Web*[3]) vorangetrieben und ihre Verwendung gefördert wird. Allerdings besteht auf Grund der unterschiedlichen Interessen aller beteiligten Partner (u.a. Bibliotheken, Dokumentationsstellen, Datenbankproduzenten, ‚freie' Anbieter von Dokumentsammlungen und Datenbanken) kaum die Aussicht, dass sich durch diese Standardisierung semantische Heterogenität restlos beseitigen lässt. [Krause 2003] Insbesondere ist eine einheitliche Verwendung von Vokabularen und Ontologien nicht in Sicht.

Im Projekt *CARMEN*[4] wurde unter anderem das Problem der semantischen Heterogenität einerseits durch die automatische Extraktion von Metadaten aus Internetdokumenten und andererseits durch Systeme zur Transformation von Anfragen über Cross-Konkordanzen und statistisch erzeugte Relationen angegangen. [Hellweg et al. 2001] Ein Teil der Ergebnisse der Arbeiten am IZ Sozialwissenschaften[5] waren statistische Relationen zwischen Deskriptoren, die mittels Kookurrenzbeziehungen berechnet wurden. Diese Relationen wurden dann für die Übersetzung von Anfragen genutzt, um zwischen verschiedenen Ontologien oder auch Freitexttermen zu vermitteln (siehe

[1] http:// dublincore.org/
[2] http://www.w3.org/RDF/
[3] http://www.w3.org/2001/sw/
[4] Sonderfördermaßnahme im Rahmen von Global-Info (*C*ontent *A*nalysis, *R*etrieval and *M*etaData: *E*ffective *N*etworking), www.mathematik.uni-osnabrueck.de/projects/carmen
[5] http://www.gesis.org/iz/

Abbildung 1). Das Ziel dieser Übersetzung ist die Verbesserung des (automatischen) Überstiegs zwischen unterschiedlich erschlossenen Dokumentbeständen, z.B. Fachdatenbanken und Internetdokumenten, als Lösungsansatz zur Behandlung semantischer Heterogenität.

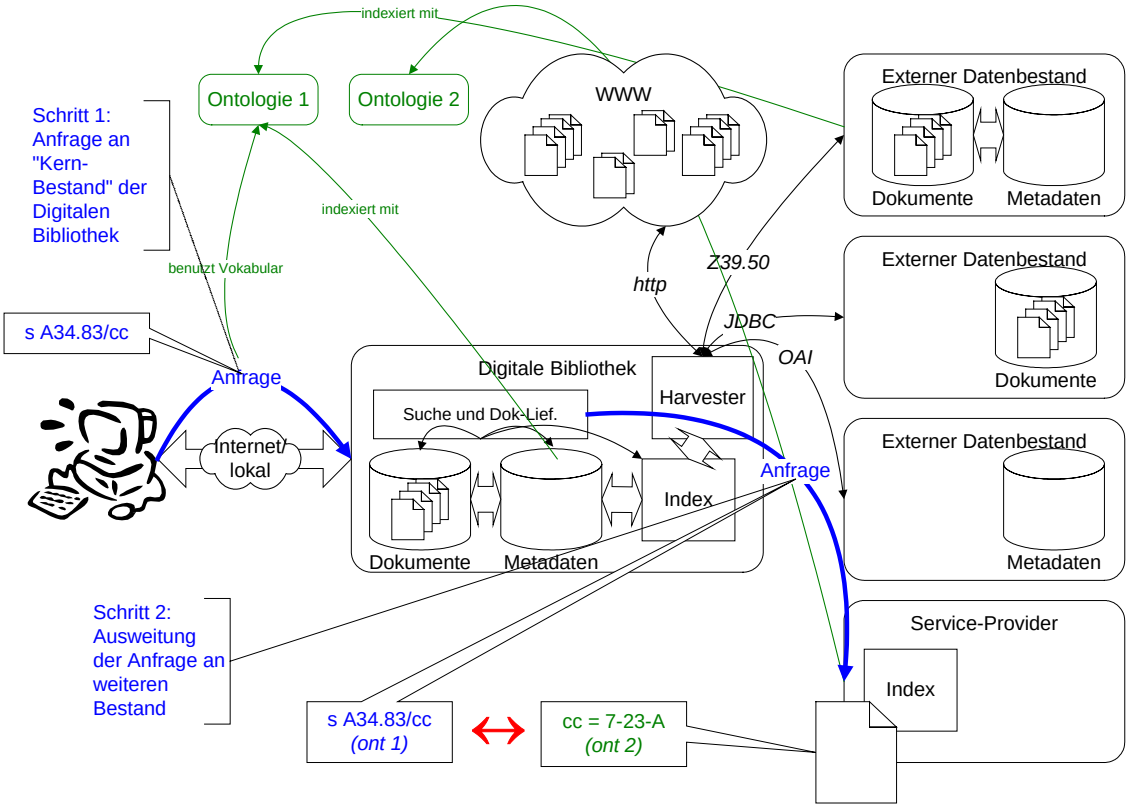

Abbildung 1: Beispiel-Szenario für semantische Heterogenität in digitalen Bibliotheken

In der Evaluierung der Verfahren zeigte sich, dass die Übersetzung von Anfragen mit Hilfe statistischer Relationen prinzipiell zu einer Verbesserung der Retrievalqualität führen konnte. Insbesondere wurde eine Verbesserung des Recall erreicht, der Ballast wurde dabei mehr oder weniger stark vergrößert. Eine systematische Analyse der Bedingungen für erfolgreiche Anwendungen konnte jedoch noch nicht geleistet werden. [Strötgen 2002a]

Es wurden allerdings einige Ansatzpunkte aufgezeigt, wie sich das Verfahren noch verbessern lassen könnte. Ein Punkt war die Verbesserung der Erzeugung oder Simulation von Doppelkorpora und die Verbindung zu Freitexttermen. An dieser Stelle setzt das Projekt *ASEMOS*[6] an, in dem die in *CARMEN* entwickelten Ansätze weiter verfolgt werden. Verbesserungen sollen durch den Einsatz von Methoden des maschinellen Lernens sowie

[6] *A*pplied *S*tatistics, *E*valuation and *M*achine Learning: *O*ntology *S*witching

linguistischer Verfahren erzielt werden, wie im Folgenden genauer dargestellt.

Vergleichbare Ansätze wurden schon im Projekt AIR/PHYS genutzt, hier allerdings für semantische Relationen zwischen Deskriptoren. [Biebricher et al. 1988] Auch das Wortschatz-Projekt der Universität Leipzig[7] berechnet auf ähnliche Weise semantische Relationen innerhalb einer Sprache. [Quasthoff 1998] Ebenso erzeugen die „EuroSpider"-Systeme[8] für multilinguales Retrieval semantische Relationen durch Kookurrenzanalysen, hier aber zwischen verschiedenen Sprachen. [Braschler & Schäuble 2000] Ähnliche semantische Analysen werden beim „Interspace"-Projekt[9] genutzt, um Konzepte automatisch aus Dokumenten zu extrahieren und Anfragen zu erweitern [Schatz et al. 1996, Chang & Schatz 1999, Chung et al. 1999, ähnlich auch bei Xu & Croft 1996].

Im Kontext des *Semantic Web* wird in einer Vielzahl von Projekten versucht, Ontologien automatisch lernen zu lassen. [Maedche 2002] Auch ein Mapping zwischen Ontologien wird hier angestrebt, auch hier werden statistische Relationen zwischen Klassen berechnet. [Doan et al. 2003]

Das Besondere an *CARMEN* und den darauf aufbauenden Arbeiten ist die statistische Erzeugung von semantischen Relationen zwischen Ontologien und Freitexttermen und deren Nutzung zur Anfrageübersetzung.

Die Relevanz der Problematik hat sich auch nach Abschluss der Arbeiten am Projekt *CARMEN* zunehmend verdeutlicht. Integrierte Portale im Kontext Digitaler Bibliotheken wie z.B. die Wissenschaftsportale Infoconnex[10] und Vascoda[11] planen den Einsatz entsprechender Anfrageübersetzungstechniken oder haben sie bereits eingeführt. Auch im Kontext des *Semantic Web* wird die Vermittlung zwischen verschiedenen Ontologien als ein zentrales Problem der Integration heterogener Dokumente angesehen und bearbeitet. Im näheren Kontext dieser Arbeit wird an der Universität Hildesheim beim virtuellen Bibliotheksregal MyShelf eine Vermittlung zwischen Klassifikationen durch Ontology Switching angewendet. [Kölle et al. 2004]

[7] http://wortschatz.uni-leipzig.de/
[8] http://www.eurospider.ch/
[9] http://www.canis.uiuc.edu/projects/interspace/
[10] http://www.infoconnex.de/
[11] http://www.vascoda.de/

2 Korpusanalyse, semantische Relationen und maschinelles Lernen

Die Berechnung von semantischen Relationen in *CARMEN* beruht auf der bedingten Wahrscheinlichkeit (siehe Abbildung 2) und dem Äquivalenzindex (siehe Abbildung 3), zwei für die Analyse von Kookurrenzen bewährten Maßen.

$$P(a \rightarrow b) = \frac{P(a \,\&\, b)}{P(b)} = \frac{\frac{C_{a\&b}}{C_{all}}}{\frac{C_b}{C_{all}}} = \frac{C_{a\&b}}{C_b}$$

Abbildung 2: Bedingte Wahrscheinlichkeit

$$P(a \leftrightarrow b) = E_{ab} = \frac{C_{ab}^{\ 2}}{C_a * C_b}$$

Abbildung 3: Äquivalenzindex

In beiden Maßen wird die Wahrscheinlichkeit berechnet, dass ein Dokument mit zwei Termen *a* und *b* gemeinsam indexiert wird.

Voraussetzung für die Berechnung beider Maße ist das Vorhandensein eines Parallelkorpus, in dem dieselben Dokumente mit zwei Ontologien erschlossen wurden. Für die Generierung der statistischen Relationen zwischen Termen wurde das Werkzeug *JESTER* benutzt. [Hellweg 2002] Das Ergebnis waren Term-Term-Matrizen, die später für die Übersetzung von Anfragen genutzt wurden. [Strötgen 2002b]

In *CARMEN* war die Besonderheit, dass eine Übersetzung von einer Ontologie zu Freitexttermen aus Internet-Dokumenten ermöglicht werden sollte. Insofern wurden Freitextterme hier wie eine Ontologie behandelt, was im Gegensatz zu einer ‚echten' Ontologie mit kontrolliertem Vokabular eine linguistische Vorverarbeitung erforderlich macht. Bei der Ontologie, zu der semantische Relationen erzeugt werden sollten, handelte es sich um den „Thesaurus Sozialwissenschaften", einen Thesaurus mit etwa 7.400 Deskriptoren und 3.700 Nicht-Deskriptoren aus dem Bereich der Sozialwissenschaften.[12]

Während für die Disziplinen Mathematik und Physik ein geeignetes Parallelkorpus bereitgestellt werden konnte, fehlte ein solcher für die

[12] http://www.gesis.org/Information/Rechercheunterst/

Sozialwissenschaften, da inhaltlich erschlossene, mit Metadaten versehene Internetdokumente im Bereich der Sozialwissenschaften kaum verfügbar waren. Daher wurde ein Parallelkorpus simuliert, indem mit der probabilistischen Suchmaschine *Fulcrum Search Server*[13] ein Trainingskorpus von unerschlossenen Internetdokumenten indexiert wurde. Die Ranking-Werte der Suchmaschine wurden dann dazu genutzt, um Deskriptoren im Sinne einer automatischen Klassifizierung gewichtet Dokumenten zuzuordnen (siehe Abbildung 4).

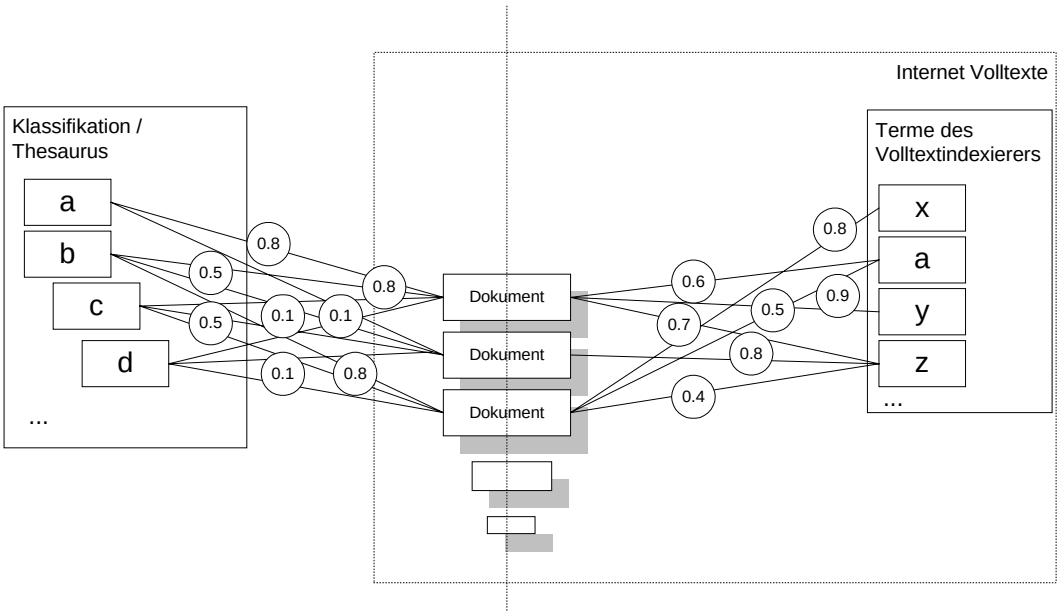

Abbildung 4: Parallelkorpus-Simulation mit vagen Deskriptoren und Volltexttermen

Diese Methode hat ganz offensichtlich ihre Schwächen, da Deskriptoren nur den Dokumenten zugeordnet werden können, die diese Begriffe mehr oder weniger wörtlich im Text enthalten. Daher werden in *ASEMOS* zur Verbesserung der Parallelkorpus-Simulation Methoden des maschinellen Lernens und der automatischen Klassifizierung eingesetzt. Der konkrete Einsatzzweck ist hier die automatische Klassifizierung inhaltlich nicht erschlossener Internet-Dokumente mit Deskriptoren aus dem Thesaurus Sozialwissenschaften. Für die Experimente wurden die Java-Bibliotheken für *Data Mining* des Projekts *WEKA*[14] eingesetzt. [Witten & Frank 2001] Die Klassifizierer wurden mit Hilfe des *GIRT*-Korpus[15] trainiert. Dieser Korpus enthält etwa 13.000 sozialwissenschaftliche bibliografische Datensätze aus den Datenbanken *SOLIS* und *FORIS*, die intellektuell mit Deskriptoren aus

[13] http://www.hummingbird.com/
[14] http://www.cs.waikato.ac.nz/ml/weka/
[15] http://www.gesis.org/Forschung/Informationstechnologie/CLEF-DELOS.htm

dem Thesaurus Sozialwissenschaften erschlossen wurden und mit Titel und Abstract auch über Freitext in einem nennenswerten Umfang verfügen.

Angesichts der großen Zahl von Freitexttermen stößt die automatische Klassifizierung auf statistischer Grundlage schnell an Grenzen. Zur linguistischen Vorverarbeitung wurden daher die extrahierten Freitextterme zunächst mit dem für die deutsche Sprache optimierten Stemmer aus dem Projekt *Apache Lucene*[16] analysiert. Die dadurch erreichte Reduzierung der Anzahl an Termen war aber bei weitem nicht ausreichend. Daher wurde auf den *Part of Speech* (POS) Tagger *QTag*[17] zurückgegriffen, um die Menge der „gestemmten" Terme auf bestimmte (besonders sinntragende) Wortklassen[18] zu reduzieren. Der POS-Tagger war vorher mit Hilfe des *NEGRA*-Korpus[19] der Universität Saarbrücken trainiert worden, einem syntaktisch annotierten Korpus mit Texten aus deutschsprachigen Zeitungen mit über 20.000 Sätzen und 350.000 Tokens. Die Textgattung Zeitung erschien als besonders geeignet für die Disziplin Sozialwissenschaften.

Wegen des hohen Aufwands bei der Berechnung statistischer Klassifikatoren und der beschränkten Ressourcen konnte nur ein kleiner Teil der *GIRT*-Dokumente zum Trainieren der Klassifizierer genutzt werden. [20] Der Ressourcenbedarf wird insbesondere durch das *Multilabel-Problem* [Sebastiani '02] verstärkt. Für die zu klassifizierenden Dokumente ist nicht nur eine Ausprägung der Zielklasse möglich, sondern mehrere; mit anderen Worten: Einem Dokument können mehrere Deskriptoren des Thesaurus zugeordnet werden. In WEKA können Multilabel-Probleme nicht direkt behandelt werden. Für jede mögliche Ausprägung der Zielklasse muss daher jeweils ein eigenen Klassifizierer trainiert werden. Für die behandelten Beispiele erhöht sich der Rechenaufwand dadurch um den Faktor 500 bis 1000.

In den durchgeführten Experimenten wurden verschiedene Arten von Klassifizierern getestet, u.a. instanzbasierte Klassifizierer sowie Support-Vector-Machine- und Naive-Bayes-Klassifizierer. [Witten & Frank 2001]

[16] http://jakarta.apache.org/lucene/
[17] http://web.bham.ac.uk/O.Mason/software/tagger/
[18] Z.B. Nomen, Adjektive und Adverbien, finite und infinite Vollverben.
[19] http://www.coli.uni-sb.de/sfb378/negra-corpus/
[20] Auf einem PC mit einem Pentium IV Prozessor (2,7 GHz Takt) und 2 GByte RAM würde die Berechnung aller Klassifizierer auf der Grundlage von nur 1000 Dokumenten ca. 80 Tage dauern.

Außerdem wurden verschiedene Arten der Gewichtung von Freitexttermen (einfaches Vorkommen, Termfrequenz und TFIDF) getestet.

Insbesondere instanzbasierte Klassifizierer lieferten viel versprechende Ergebnisse, auf Grund des besonders hohen Ressourcenbedarfs dieser Verfahren konnten aber mit bis zu 500 nur sehr wenige Trainingsdokumente genutzt werden. Ähnliches gilt für die Klassifizierer, die auf Support-Vector-Machine basieren. Naive-Bayes-Klassifizierer erlaubten ein wesentlich umfangreicheres Training, erzielten aber trotzdem bislang nicht hinreichende Ergebnisse bei den anschließenden Evaluierungen.

Die trainierten Klassifizierer wurden dann genutzt, um etwa 7.000 Internet-Dokumente (ausschließlich HTML-Dokumente) aus dem *CARMEN*-Testkorpus mit sozialwissenschaftlichen Texten zu klassifizieren. Dabei wurden die HTML-Dokumente mit Hilfe von *JTidy*[21] bereinigt und in DOM-Dokumente konvertiert. Abschließend wurden mit *XPath*-Anfragen und der XML-Bibliothek *Apache Xalan*[22] die Text-Knoten aus dem Titel und dem Textkörper extrahiert und für die Klassifizierung genutzt. Dabei wurde die Annahme zu Grunde gelegt, dass Dokumente aus den Beständen *GIRT* und *CARMEN* hinreichend ähnlich sind, um brauchbare Ergebnisse zu erzielen.

Eine weitere wesentliche Schwäche der *CARMEN*-Implementierung bestand darin, dass die Zahl der berechneten Freitextterme ausgesprochen beschränkt war. Als Freitextterme wurden nur die Begriffe genutzt, die auch im Thesaurus Sozialwissenschaften als Deskriptoren oder Nicht-Deskriptoren verzeichnet sind. Bei der späteren Nutzung für Anfrageübersetzungen hat dies natürlich insbesondere dann sehr negative Auswirkungen, wenn von Freitexttermen aus übersetzt werden soll, aber kaum ‚echte' Freitextterme für eine Übersetzung zur Verfügung stehen. Als Konsequenz waren in *CARMEN* nur Übersetzungen vom Thesaurus nach Freitexttermen hin evaluiert worden.

In *ASEMOS* wurde die Zahl der Freitextterme dagegen deutlich erhöht. Dabei wurde die oben beschriebenen Stemmer und POS-Tagger genutzt, um über eine hohe Anzahl gestemmter Terme die Wahrscheinlichkeit einer möglichen Übersetzung einer Anfrage drastisch zu verbessern.

[21] http:/jtidy.sourceforge.net/
[22] http://xml.apache.org/xalan-j/

3 Erste Ergebnisse

Zur Evaluation der Qualität der mit *WEKA* erzeugten und dem *GIRT*-Korpus trainierten automatischen Klassifikatoren wurden *GIRT*-Dokumente automatisch klassifiziert und die Ergebnisse mit den intellektuell zugeordneten Deskriptoren verglichen. Die Klassifizierer waren mit einer Zufallsstichprobe von 1.000 bis 5.000 Dokumenten trainiert worden, für die Evaluierung wurde eine zweite Zufallsstichprobe im Umfang von 100 *GIRT*-Dokumenten erstellt.

Für jeden Klassifizierer wurden für die Dokumente richtig und falsch klassifizierte Deskriptoren verglichen, als Kriterium galt dabei die intellektuelle Zuordnung in den *GIRT*-Dokumenten (siehe Abbildung 5).

Klassifizierer	**Typ**	**Train.-Dok.**	**Deskrip-toren**	**true pos.**	**false pos.**	**false neg.**	**t/f pos.**
NaiveBayes	nom.	1.000	5,75	0,56	5,19	5,77	13,00 %
NaiveBayes	nom.	5.000	0,48	0,04	0,44	8,25	1,21 %
NaiveBayes	Tf	1.000	0,64	0,46	0,18	5,71	25,00 %
NaiveBayes	Tfidf	1.000	266,05	4,00	262,05	2,03	1,62 %
NaiveBayes	tfidf + discr.	1.000	5,28	1,80	3,48	4,93	68,03 %
NaiveBayes	tfidf + kernel	1.000	79,44	1,99	77,45	4,28	6,59 %
NaiveBayes	Tfidf	5.000	0,05	0,06	1,45	8,17	4,44 %
NaiveBayes	tfidf + discr.	5.000	11,31	2,99	8,32	4,26	57,3 %

Abbildung 5: Mittelwerte der Ergebnisse der automatischen Klassifizierer

Die Ergebnisse dieser Methode sind insofern mit Vorsicht zu interpretieren, als dass eine intellektuelle Zuordnung von Deskriptoren in der Regel kaum erschöpfend ist. Daher ist es durchaus wahrscheinlich, dass vom automatischen Klassifizierer ‚falsch' zugeordnete Deskriptoren tatsächlich zum Inhalt eines Dokuments passen. Beispielsweise war ein Dokument unter anderem intellektuell mit den Deskriptoren „China" und „Interkulturelle Kommunikation" erschlossen worden. Der Klassifizierer ordnete unter anderem die Deskriptoren „Asien", „Ost-Asien" und „Kultur" zu, die alle als *false positives* zählen, obwohl sie zum Dokument passen. Bei der intellektuellen Überprüfung einer Stichprobe hat sich gezeigt, dass z.B. für den Klassifizierer Naive Bayes mit Diskretisierungs-Option ungefähr die Hälfte aller *false positives* eigentlich als *true positives* gezählt werden müssten. Das erklärt zumindest teilweise die große Zahl von Fehlern in

Abbildung 5 und schränkt die Aussagekraft der automatischen Auswertung ein.

Einige der trainierten Klassifizierer wurden anschließend verwendet, um die Internet-Dokumente aus dem *CARMEN*-Korpus zu klassifizieren. Das Ergebnis wurde mit dem des oben beschriebenen *CARMEN*-Verfahrens verglichen (siehe Abbildung 6).

Klassifizierer	**Typ**	**Ø Deskriptoren**	**Standardabweichung**	**Überschneidung mit *CARMEN***
CARMEN		39,75	71,29	./.
ASEMOS NaiveBayes	tfidf + discr.	7,91	13,84	3,05
ASEMOS NaiveBayes	tfidf + kernel	15,68	25,25	0,61

Abbildung 6: Vergleich *ASEMOS*-Klassifizierer mit *CARMEN*-Klassifizierern

Dabei zeigte sich, dass die Naive-Bayes-Klassifizierer deutlich weniger Deskriptoren zuordneten als der *CARMEN*-Klassifizierer. In *CARMEN* war eines der Probleme, dass teilweise viel zu viele Deskriptoren einem Dokument zugeordnet wurden und daher die Anfrageübersetzung zu sehr viel Ballast führte. Dieses Problem könnte mit den neuen Klassifizieren evtl. gelöst werden. Allerdings ist natürlich nicht allein die Anzahl, sondern vor allem die Qualität der Zuordnungen entscheidend. Hier zeigt sich, dass der Klassifizierer mit Diskretisierungs-Option zwar weniger Deskriptoren zuordnete als der mit Kernel-Abschätzung, dass aber die Überschneidung mit den *CARMEN*-Zuordnungen deutlich größer war. Für eine gesicherte Bewertung ist es nun erforderlich, die Zuordnungen intellektuell zu überprüfen und die Auswirkungen auf die Retrievalqualität zu evaluieren.

Schließlich wurden basierend auf den Ergebnissen der neuen Klassifikatoren neue Doppelkorpora simuliert, die dann Grundlage für eine neue Berechnung von Term-Term-Matrizen waren. Ein Vergleich ist auch hier sehr schwierig, weil in *CARMEN* die Zuordnungen nur zu „Pseudo-Freitexttermen“ erfolgte, während in *ASEMOS* gestemmte und über Wortklassen und Schwellenwerte selektierte ‚echte’ Freitextterme verwendet wurden. Die Zahl der einem Thesaurus-Term zugeordneten Freitextterme ist erwartungsgemäß deutlich höher. Die Evaluierung der Auswirkungen auf das Retrievalergebnis steht allerdings noch aus. Auch diese ist nur bedingt vergleichbar, da eine Übersetzung von Freitexttermen in Thesaurus-Terme, wie oben beschrieben, mit den *CARMEN*-Ergebnissen nicht sinnvoll durchführbar ist.

4 Zusammenfassung und Ausblick

Auch wenn bisher gezeigt werden konnte, dass der Einsatz von Methoden des maschinellen Lernens und eine ausgefeiltere linguistische Behandlung von Freitexttermen deutliche Auswirkungen auf die Erstellung semantischer Relationen haben, ist die Auswirkung auf die Retrievalqualität bislang erst in Ansätzen erkennbar. Erste Ergebnisse lassen die Vermutung zu, dass die Zuordnung einer geringeren Anzahl von Deskriptoren zu Dokumenten eine leichte Verringerung des Ballastes bewirken könnten. Die tatsächlichen Auswirkungen auf die Retrievalqualität müssen aber nun empirisch überprüft werden.

Beim Einsatz automatischer Klassifizierer besteht vermutlich noch einiges Potenzial zur Optimierung der Parameter. Solange die Effekte der Doppelkorpussimulation nicht klar sind, ist dies aber nur in enger Koppelung mit der empirischen Überprüfung der Auswirkungen auf die Retrievalqualität möglich. Erschwerend kommt hinzu, dass das Training der Klassifikatoren hohe Anforderungen an die Ressourcen stellt und trotzdem oft mehrere Tage Rechenzeit erfordert. Dies verlangsamt die Optimierungs- und Anpassungszyklen erheblich.

Eine neue Chance zur Verbesserung der Qualität hat sich inzwischen durch das neue *SozioNet*-Korpus[23] der TU Darmstadt ergeben. In diesem Korpus sind einige Hundert sozialwissenschaftliche Internet-Dokumente intellektuell mit Deskriptoren des Thesaurus Sozialwissenschaften indexiert worden. Mit der Nutzung dieses Korpus wäre eine Simulation des Parallelkorpus evtl. nicht mehr nötig, die Vagheit der Grundlage für die Term-Term-Matrizen könnte dadurch zumindest reduziert werden.

In einem weiteren Schritt sollen die auf Grundlage der sozialwissenschaftlichen Dokumente entwickelten Methoden in Kooperation mit dem FIZ Karlsruhe[24] auf die Domäne Patentdokumentation übertragen und auf ihre Wirksamkeit hin evaluiert werden.

5 Literaturverzeichnis

[Biebricher et al. 1988] Biebricher, Peter ; Fuhr, Norbert ; Lustig, Gerhard ; Schwantner, Michael ; Knorz, Gerhard: The automatic indexing system AIR/PHYS - from research

[23] http://www.sozionet.org/
[24] http://www.fiz-karlsruhe.de/

to applications. In: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, 1988, S. 333–342.

[Braschler & Schäuble 2000] Braschler, Martin ; Schäuble, Peter: Using Corpus-Based Approaches in a System for Multilingual Information Retrieval. In: Information Retrieval 3 (2000), Nr. 3, S. 273–284

[Chang & Schatz 1999] Chang, Conrad T. K. ; Schatz, Bruce R.: Performance and implications of semantic indexing in a distributed environment. In: Proceedings of the eighth international conference on Information and knowledge management. ACM Press, 1999, S. 391–398

[Chung et al. 1999] Chung, Yi-Ming ; He, Qin ; Powell, Kevin ; Schatz, Bruce R.: Semantic indexing for a complete subject discipline. In: Proceedings of the fourth ACM conference on Digital libraries. ACM Press, 1999, S. 39–48

[Doan et al. 2003] Doan, AnHai; Madhavan, Jayant ; Dhamankar, Robin ; Domingos, Pedro ; Halevy, Alon: Learning to match ontologies on the Semantic Web. In: The VLDB Journal 12 (2003), Nr. 4, S. 303–319.

[Hellweg 2002] Hellweg, Heiko : Einsatz von statistisch erstellten Transferbeziehungen zur Anfrage-Transformation in ELVIRA. In: Krause, Jürgen ; Stempfhuber, Max (Hrsg.): Integriertes Retrieval in heterogenen Daten. Text-Fakten-Integration am Beispiel des Verbandinformationssystems ELVIRA. (Forschungsberichte des IZ Sozialwissenschaften) Bonn: IZ Sozialwissenschaften, 2002.

[Hellweg et al. 2001] Hellweg, Heiko; Krause, Jürgen ; Mandl, Thomas ; Marx, Jutta ; Müller, Matthias N. ; Mutschke, Peter ; Strötgen, Robert: Treatment of Semantic Heterogeneity in Information Retrieval. Bonn : IZ Sozialwissenschaften, 2001 (IZ-Arbeitsbericht; Nr. 23)

[Kölle et al. 2004] Kölle, Ralph ; Mandl, Thomas ; Schneider, René ; Strötgen, Robert: Weiterentwicklung des virtuellen Bibliotheksregals MyShelf mit Semantic Web-Technologie: Erste Erfahrungen mit informationswissenschaftlichen Inhalten. In: Schmidt, Ralph (Hrsg.): Information Professional 2011; 26. Online-Tagung der DGI, DGI, Frankfurt am Main, 15. bis 17. Juni 2004; Proceedings. Frankfurt am Main : DGI, 2004 (Tagungen der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis; 7)

[Krause 2003] Krause, Jürgen: Standardisierung von der Heterogenität her denken. Zum Entwicklungsstand bilateraler Transferkomponenten für digitale Fachbibliotheken. Bonn : IZ Sozialwissenschaften, 2003 (IZ-Arbeitsbericht; Nr. 28)

[Maedche 2002] Maedche, Alexander: Ontology Learning for the Semantic Web. Boston et al : Kluwer Academic Publishers, 2002 (Kluwer International Series in Engineering & Computer Science)

[Quasthoff 1998] Quasthoff, Uwe: Tools for Automatic Lexicon Maintenance: Acquisition, Error Correction, and the Generation of Missing Values. In: Proceedings of the first International Conference on Language Resources & Evaluation, ELRA 1998, S. 853–856

[Schatz et al. 1996] Schatz, Bruce R. ; Johnson, Eric H. ; Cochrane, Pauline A. ; Chen, Hsinchun: Interactive term suggestion for users of digital libraries: using subject thesauri and co-occurrence lists for information retrieval. In: Proceedings of the first ACM international conference on Digital libraries, ACM Press, 1996, S. 126–133.

[Sebastiani '02] Sebastiani, Fabrizio: Machine learning in automated text categori-zation. In: ACM Comput. Surv. 2002, 34, Nr. 1; S. 1--47.

[Strötgen 2002a] Strötgen, Robert: Behandlung semantischer Heterogenität durch Metadatenextraktion und Anfragetransfer. In: Womser-Hacker, Christa (Hrsg.) ; Wolff, Christian (Hrsg.) ; Hammwöhner, Rainer (Hrsg.): Information und Mobilität: Optimierung und Vermeidung von Mobilität durch Information; Proceedings des 8. Internationalen Symposiums für Informationswissenschaft (ISI 2002). Konstanz : UVK, 2002 (Schriften zur Informationswissenschaft; Bd. 40), S. 259–271.

[Strötgen 2002b] Strötgen, Robert: Fachdatenbanken und Internet-Quellen: Rechercheüberstieg durch Anfragetransfer. In: Schubert, Sigrid (Hrsg.) ; Reusch, Bernd (Hrsg.) ; Jesse, Norbert (Hrsg.): Informatik bewegt. Informatik 2002: 32. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 30. September - 3. Oktober 2002 in Dortmund; Proceedings; Ergänzungsband. Bonn : Gesellschaft für Informatik, 2002 (Lecture Notes in Informatics; P-20), S. 52–56

[Witten & Frank 2001] Witten, Ian H. ; Frank, Eibe: Data Mining. Praktische Werkzeuge und Techniken für das maschinelle Lernen. München : Hanser, 2001

[Xu & Croft 1996] Xu, Jinxi ; Croft, W. B.: Query expansion using local and global document analysis. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, 1996, S. 4–11