INSTITUT FÜR DEUTSCHE SPRACHE

OPAL

Online publizierte Arbeiten zur Linguistik

ISSN 1860-9422

1/2005

Christian Weiß

Die thematische Erschließung von Sprachkorpora

OPAL – Online publizierte Arbeiten zur Linguistik
Herausgegeben vom Institut für Deutsche Sprache

Institut für Deutsche Sprache
Postfach 10 16 21
68016 Mannheim
opal@ids-mannheim.de

Technische Redaktion: Norbert Volz

Christian Weiß

Die thematische Erschließung von Sprachkorpora[1]

Zusammenfassung

Ziel des Teilprojekts[2] ist die thematische Erschließung der Korpora, um sowohl themenspezifische virtuelle Subkorpora zusammenstellen zu können als auch aufgrund der Analyse sachgebietsbezogener Häufigkeitsverteilungen z.B. Lesarten disambiguieren zu können. Ausgangspunkt ist die Erstellung einer Taxonomie von Sachgebietsthemen. Dies erfolgt in einem semiautomatischen Verfahren, welches die Anwendung von *Textmining* (*Dokumentclustering*) und die manuelle Zuordnung von Clustern in eine externen Ontologie beinhaltet. Es wird argumentiert, dass die so gewonnene Taxonomie sowohl intuitiver als auch objektiver ist als bestehende, rein manuelle Ansätze. Sie eignet sich zudem gleichermaßen für manuelle als auch für maschinelle Klassifikation. Für letzteres wird der *Naive Bayes'sche Textklassifikator* motiviert und für ein klassifiziertes Korpus von knapp zwei Milliarden Wörtern evaluiert.

1. Motivation

Eine zentrale Aufgabe der Korpuslinguistik ist die Identifizierung von Varianzen beispielsweise zur Lesartendisambiguierung eines Lexems wie etwa „Maus“. Die bloße elektronische Verfügbarkeit von Sprachdaten wäre schon bei diesem einfachen Anwendungsbeispiel unzureichend und müsste durch die Festlegung von Subkorpora wie „Zoologie“ oder „EDV“ ergänzt werden. Ein Ansatz, der eine Katalogisierung von Sprachkorpora vornimmt, ist der von Kučera/Francis (1979). Die Unterteilung der Kategorien ist jedoch hier in der Regel nicht an inhaltliche, sondern an formale, textgattungsspezifische Kriterien geknüpft und daher für eine Anwendung wie die oben skizzierte zu grobkörnig. Eine feingliedrigeres, inhaltsbasiertes Klassifikationsschema ist das von PAROLE (2004). Hier sind beispielsweise die oben angesprochenen Kategorien „Zoologie“ und „EDV“ vorhanden. Bezüglich des Klassifikationsschemas können jedoch zwei Anmerkungen gemacht werden. Zum einen stellt sich die Frage nach Vollständigkeit: Gibt es nicht vielleicht Kategorien, die relevant sein könnten, jedoch nicht genannt werden? So drängt sich bei der Existenz einer Kategorie „Europäische Union“ fast zwingend die Frage auf, warum nicht vergleichbare Organisationen genannt werden. Die zweite Anmerkung betrifft die Tatsache, dass Zeitungsartikel, aus denen ja der Großteil eines Korpus besteht, jedenfalls das des IDS, nicht klassifiziert werden.

Beides, sowohl ein möglichst allumfassendes Klassifikationsschema und die Klassifizierbarkeit jedes Dokumentes eines Korpus ist Gegenstand des am Institut für Deutsche Sprache durchgeführten Teilprojekts „Thematische Erschließung“. Bezüglich der erstrebten Vollständigkeit eines Klassifikationsschemas werden zwei Komponenten vorgeschlagen: zum einen der Einsatz von Textminingverfahren bzw. eines Dokumentclusterers, der Indikatoren für die thematische Zusammensetzung von Zeitungskorpora liefert, zum anderen die Verwendung einer höherstufigen Ontologie wie die des *Open Directory*-Projekts, welche – wenigstens vom Anspruch her – ein allumfassendes und normiertes Themeninventar liefert. Das zweite Ziel, nämlich die Klassifizierbarkeit des Gesamtkorpus, also aller Texte einschließlich Texten aus Zeitungskorpora, erfolgt in Form von vollautomatisierter Textklassifikation, wobei ich im Sinne einer möglichst hohen Robustheit für den *Naiven Bayes'schen Textklassifikator* argumentiere.

---

[1] Kommentare bitte an `weiss@ids-mannheim.de`. Cyril Belica, Rainer Perkuhn, Marc Kupietz und Norbert Volz danke ich für wertvolle Anmerkungen.

[2] Projekt: „Methoden der Korpusanalyse und -erschließung“, Arbeitsgruppe „Korpustechnologie“

Der Rest des Artikels gliedert sich wie folgt: Im nächsten Kapitel stelle ich den Aufbau des Teilprojekts und dessen einzelne Komponenten vor. Eine Evaluierung des Systems erfolgt in Kapitel 3. Weitere Anwendungen werden in Kapitel 4 angesprochen. Eine Zusammenfassung erfolgt in Kapitel 5.

2. Aufbau

In diesem Kapitel möchte ich das erarbeitete Verfahren vorstellen. Es ist in Abb. 1 schematisiert und gliedert sich auf oberster Ebene in zwei Schritte. Der erste umfasst die Arbeitsschritte, die durch die gepunkteten Pfeile dargestellt sind. Er beinhaltet die Konstruktion einer *Thementaxonomie* bzw. eines Trainingskorpus durch einen (menschlichen) Annotierer, der jedoch Vorgaben erhält: Zum einen in Form von *Clustern*, die durch einen *Dokumentclusterer* erzeugt werden, zum anderen in Form einer externen Ontologie wie der des *Open Directory*-Projekts. Im zweiten Schritt erfolgt die automatische Klassifikation von Korpustexten unter Zuhilfenahme der Trainingsdaten. Ich werde im Folgenden die Komponenten im Einzelnen beschreiben.

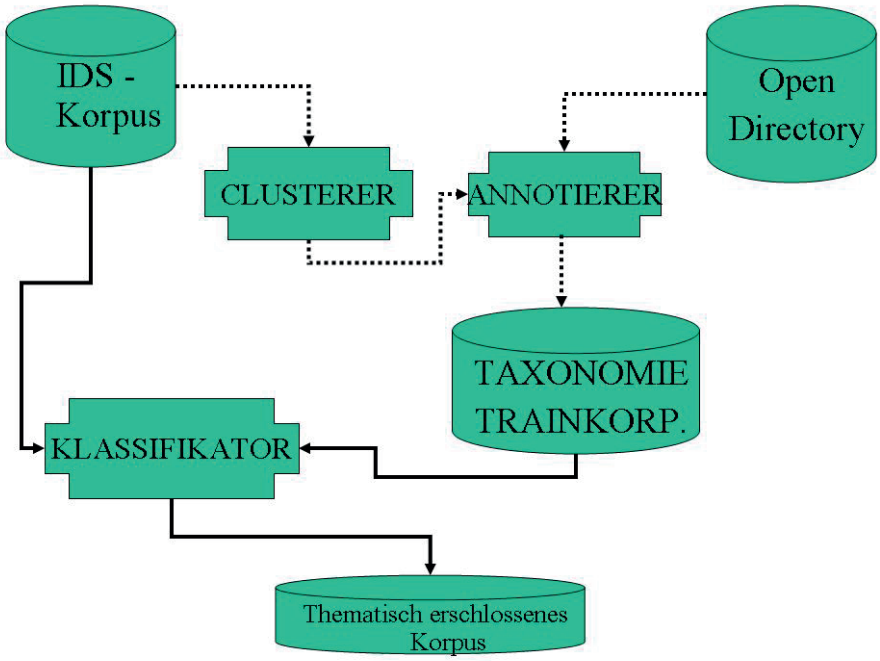

Abb. 1: Übersicht über die Bestandteile des Teilprojekts

2.1 Dokumentclustering

Clustering ist eine Unterform des *Data Minings* bzw. *unüberwachten maschinellen Lernens* und wird hauptsächlich zur *explorativen Datenanalyse* eingesetzt z.B. in der Biologie, den Sozialwissenschaften oder im Information Retrieval.[3] *Dokumentclustering* bedeutet die automatische Gruppierung von Texten mit ähnlichem Inhalt. Die formale Grundlage ist hierbei die so genannte *bag of words*-Annahme, nach der die thematische Ähnlichkeit zweier Dokumente durch die Anzahl der gemeinsamen Wörter bzw. *Terme* messbar ist.[4] Da jedoch nicht alle Terme von gleichem Gewicht sind, ist ein Verfahren nötig, das die *Termgewichtung* berechnet. Hier wurde das in Spark-Jones (1972) beschriebene *idf*-Gewichtungsschema verwendet. Eine weitere wichtige Verfeinerung besteht in der Berücksichtigung unterschiedlicher Doku-

[3] Siehe Jain/Dubes (1988).

[4] Da Deutsch eine Sprache mit einem reichen morphologischen Inventar ist, wurde ein Lemmatisierer eingesetzt. Siehe auch Belica (1994).

mentlängen durch die *Vektorkosinusfunktion* im Sinne von Salton (1968). Wegen der Normierung der Vektorlänge fallen unterschiedliche Dokumentlängen weniger ins Gewicht. Dies ist besonders wichtig bei Korpora, die aus Zeitungstexten zusammengesetzt sind.

Hinsichtlich eines geeigneten *Clusteralgorithmus* standen zwei Grundtypen zur Wahl: *Nicht hierarchisches Clustering* oder *hierarchisches Clustering*. Bei nicht hierarchischen Ansätzen wie z.B. dem *k-means-Verfahren* müssen sowohl die Anzahl der Cluster als auch deren Inhalt in Form einer extern klassifizierten Datenuntermenge bzw. *Seeds* schon vor der Laufzeit bekannt sein. Hierarchische Ansätze erzeugen (rekursiv oder induktiv) Baumstrukturen bzw. *Dendrogramme*, wobei der oberste Knoten die gesamte Dokumentmenge darstellt und die Blätter einzelne Dokumente. Hierarchische Ansätze benötigen keine vorklassifizierten Daten, sind jedoch bedeutend langsamer als nichthierarchische. Darüber hinaus kommt es häufig zu einer sehr großen Anzahl von Knoten und wegen asymmetrischer Füllungen mit dominierten Dokumenten zu unintuitiven Baumstrukturen. In der Praxis werden daher oft Mischverfahren verwendet wie beispielsweise der *Buckshot-Algorithmus* oder das hier verwendete und auf Karypis/Zhao (2002) zurückgehende Verfahren: Hierbei wird wie beim hierarchischen Clustering eine Baumstruktur erzeugt, jedoch nur mit einer Untermenge der Daten und mit einer vorgegeben Anzahl von *k* Clustern. Es folgt eine globale Optimierung dieses Zwischenergebnisses und die Zuweisung der Restdokumente. Da das Verfahren top-down funktioniert, Knoten also rekursiv eingebettet sind und das Ergebnis global bewertet wird, ist die Angabe einer festen Größe *k* weniger problematisch als es anfangs scheint. So sind Dokumente, die ein *k+i*-tes Thema repräsentieren könnten wie z.B. „Skifahren“ unter dem global geeignetsten der *k* Knoten eingebettet, wie etwa dem Knoten, der „Wintersportdokumente“ dominiert.[5] Eine weitere Anforderung an den Clusterer ist die Quantifizierbarkeit thematischer Nähe durch *Routing*, d.h., es soll unterschieden werden können, inwiefern ein Dokument einen eher typischen Repräsentant eines Clusters darstellt oder nicht. Maßgeblich war hierfür das *z-score-Maß* bzw. die Distanz eines Dokuments zu einem Clusterzentrum bzw. *Zentroiden*.

2.1.1 Die Gewinnung von Clustern

Eine Anwendung des oben skizzierten Verfahrens auf Sporttexte ist in Karypis (2002) beschrieben. Darin wird gezeigt, wie das Programm aus einem Sporttextekorpus eine selbstständige thematische Segmentierung vornimmt. Wegen des tagespolitischen Bezugs von Pressetexten und der damit verbundenen hohen Streuung der relevanten Terme und der Fülle der zu erwartenden Themen musste das Verfahren, um es zu übernehmen, leicht modifiziert werden: Es erfolgte eine jahrgangsweise Bearbeitung der Daten zu je 100 Clustern pro Jahrgang. Diese Zahl wurde nach mehreren Probedurchläufen gewählt, deren Ziel es war, mit möglichst wenig Clustern eine möglichst große Anzahl von Themen abzudecken, wobei eine leichte Redundanz (ein Thema – zwei Cluster) für ein manuelles Verfahren sich als ideale Voraussetzung darstellte. Um möglichst nur korrekte Daten zu erhalten, wurden pro Cluster nur die 300 gewichtigsten (gewichtig nach dem oben angesprochenen *z-score*-Maß) Dokumente übernommen.

Dem vollautomatischen Clustering folgten zwei manuelle Schritte. Der erste Schritt bestand in einer Qualitätskontrolle: Cluster, die keine thematische Homogenität im erwünschten Sinne aufwiesen, wurden ausgeschlossen, oder, falls es sich um seltene Themen wie beispielsweise „Reitsport“ handelte, komplett, d.h. Dokument für Dokument überprüft. Der zweite Schritt bestand in der Annotierung, bei der jedes Cluster zum einen spontan nach Inhalt, zum anderen gemäß der in der nächsten Sektion erklärten Thementaxonomie annotiert wurde. So wurde beispielsweise ein Cluster mit Texten über die Krankheit „Aids“ aus dem Jahr 1985 mit der

[5] Siehe Karypis/Zhao (2002) zur weiteren Information.

Bezeichnung „:aids_85“ und dem Themengebiet „Gesundheit_Ernaehrung: Gesundheit“ markiert. Dieses und andere Beispiele können Tabelle 1 entnommen werden.

Ausgewählte Sachthemen und deren spontane und thematische Annotierung:	
Spontane Annotierung	Annotierung nach Thema
: aids_85	Gesundheit_Ernaehrung: Gesundheit
: affaere_barschel_87	Politik: Inland
: bse_01	Gesundheit_Ernaehrung: Gesundheit
: chemieunfall_sandoz_86	Technik_Industrie: Unfaelle
: eherecht_scheidungsrecht	Staat_Gesellschaft: Recht
: elfter_september_01	Politik: Ausland
: filmkritik_02	Kultur: Film
: fussball_uefa_94	Sport: Fussball
: polizeieinsaetze_02	Staat_Gesellschaft: Verbrechen
: tierbeobachtungen_86	Natur_Umwelt: Tiere
: klassische_musik_00	Kultur: Musik
: krieg_indien_pakistan_02	Politik: Ausland
: personal_computer_85	Technik_Industrie: EDV_Elektronik
: prozess_barbie_87	Staat_Gesellschaft: Drittes_Reich_Rechtsextremismus
: reiten_92	Sport: Vermischtes
: skispringen_97	Sport: Wintersport
: wende_ddr_89	Politik: Inland
: wirtschaftsentwicklung_00	Wirtschaft_Finanzen: Sozialprodukt

Tabelle 1: Auswahl an Clustern sowie Zuordnung zu Kategorien

2.1.2 Die Gewinnung von Clusterdaten bei Themen mit geringer Häufigkeit

Das Verfahren lieferte thematisch homogene Trainingsdaten für häufig auftretende Themen. In einigen Fällen, wie beispielsweise „Kultur: Mode“ oder „Staat_Gesellschaft: Drittes_Reich _Rechtsextremismus“ war jedoch die gefundene Datenmenge zu gering oder es ließen sich keine Daten finden. In solchen Fällen erfolgten Suchanfragen, um geeignete Dokumente zu finden. Um eine möglichst umfangreiche Belegmenge zu erhalten, d.h. eine mit einem möglichst hohen *Recall*, wurde darauf geachtet, nicht zu spezifisch bei der Wahl der Stichwörter zu sein, beispielsweise durch Suche nach Wörtern mit der Teilkette „nazi“. Eine Präzisierung der Ergebnisse erfolgte dann durch Einsatz des Clusterers. Hierdurch konnten dann Texte, die zwar häufig in Verbindung mit der Suchanfrage auftraten, jedoch für das betreffende Thema nicht relevant waren, wie z.B. Dokumente über eine pakistanische Ministerpräsidentin „Benazir Bhutto“, erkannt und vom weiteren Verfahren ausgeschlossen werden. Eine weitere Filterung erfolgte dann, analog zu dem oben beschriebenen direkten Verfahren, durch *Routing*, d.h. die Anordnung der gefundenen Themen nach Gewicht, wodurch der Suchraum für die manuelle Evaluierung einschränkt werden konnte.

2.1.3 Unerwünschte Daten

Einige der Cluster enthielten pressetypische Angaben wie „Veranstaltungshinweise“, „Inhaltsverzeichnisse“, „Apothekennotdienste“, etc. Diese „Schrottdaten“ wurden für das weitere Verfahren nicht herausgefiltert, sondern für einen in Kap. 4.3 beschriebenen Filter für Datenmüll bzw. *Spam Checking* eingesetzt.

2.1.4 Zusammensetzung der Cluster

Das so entstandene Korpus umfasst 416 960 Dokumente. Der größte Teil, 300 378 Dokumente, wurde durch das erste, rein clusterbasierte Verfahren gewonnen, 37 955 davon waren Schrottdaten. Die restlichen Daten verteilen sich auf 1500 Cluster. 77 569 Dokumente konnten durch das zweite, durch eine Suchanfrage modifizierte Verfahren bestimmt werden. Für 1058 Dokumente, die aus Fortsetzungsromanen bestanden, wurde kein automatisiertes Verfahren angewendet. Die Zusammensetzung ist grafisch in Abb. 2 wiedergegeben. Die Präzision des Korpus, d.h. der relative Anteil korrekter Texte, belief sich auf 90%.

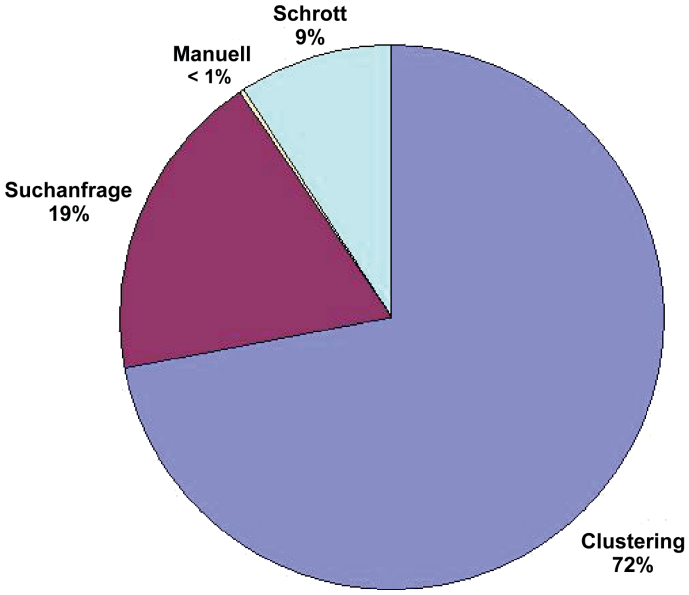

Abb. 2: Zusammensetzung der Cluster

2.2 Der Themenkatalog

Wie bereits oben angesprochen, ist ein Teilziel des Vorhabens ein Sachgebietskatalog mit einer möglichst objektiven, d.h. externen Themenbeschreibung. Da keine Beschränkung hinsichtlich der zu klassifizierten Themen erfolgen sollte, wie etwa die Beschränkung auf wissenschaftliche Themen bei einer wissenschaftlichen Bibliothek, erfolgte die Ausrichtung anhand einer höheren, möglichst allumfassenden Ontologie. Die wohl größte existierende Ontologie bildet das *Open Directory*, ein Katalog für Web-Dokumente mit 590 000 Kategorien, 66 849 Editoren und 4 Millionen manuell annotierten Dokumenten.[6] Wegen Ihres Anspruchs, alle Themengebiete zu erfassen und zu normieren, liefert das *Open Directory* einen großen Pool möglicher Themen und Themenbeschreibungen. Gegen eine direkte Übernahme als Klassifikationsschema sprachen jedoch mehrere Punkte: Zum einen erwies sich nur ein Bruchteil der Kategorien als interessant. So besteht zwar ein Interesse an der Übernahme einer Kategorie „Kultur: Film“, jedoch nicht an einer Kategorie „Kultur: Film: Filmverleih“. Ein weiterer Punkt, der gegen eine Eins-Zu-Eins-Übernahme sprach, waren „versteckte Kategorien“, d.h. eng verwandte Themen mit sehr unterschiedlichen Topkategorien. So werden beispielsweise für Gartenthemen so unterschiedliche Kategorien angegeben wie: „Zuhause: Garten“, „Wirtschaft: Bauwesen: Garten und Landschaftsbau“, „Wissenschaft: Naturwissenschaft: Biologie: Botanik: Botanische Gärten“ oder „Wirtschaft: Konsumgüter: Haus_und_Garten“. Neben einem zu hohen Maß an Feingliedrigkeit konnte jedoch auch die gegenteilige Tendenz beobachtet werden, nämlich ein zu grobes Raster, beispielsweise bei literarischen oder religiösen Texten.

[6] Stand März 2005. Siehe Netscape (2002) zur näheren Information.

Um die oben skizzierten Probleme zu vermeiden, musste eine separate Taxonomie von Themen, wie in Tabelle 2 verdeutlicht, erstellt werden, die in mehreren Punkten von der des *Open Directory* abweicht. Zum einen wurden für einen Sprachkorpus nicht relevante Kategorien herausgefiltert und die Taxonomie hierdurch erheblich reduziert, zum anderen erfolgte eine Umgewichtung: Dokumente mit Bezug zu Themen wie „Religion“, „Fiktion“ oder „Wissenschaft“ werden als Belege für eine besondere Art von Sprache aufgefasst wie „religiöse“, „literarische“ oder „wissenschaftliche“ Sprache. Die relevanten Themen sollten daher auf oberster Ebene kodiert werden. Sie entsprechen daher eher einem „Genre“ als einer thematischen Kategorie und werden in Tabelle 2 kursiv dargestellt. „Versteckte“ Kategorien wurden nicht zugelassen, sondern durch eine neue Kategorie ersetzt, beispielsweise durch die Kategorie „Natur_Umwelt: Garten“.

Neben diesen Nachteilen, die gegen eine direkte Übernahme des *Open Directory* und für die Erstellung einer eigenen Themenontologie sprachen, bleibt der Vorteil der Normierbarkeit von Themen durch eine externe Instanz bzw. die Möglichkeit einer Verankerung wie sie in Tabelle 3 ausschnittsweise dargestellt wird.[7] Auf diese Weise werden interne Themen durch externe definiert (wie z.B. durch Und-Verknüpfungen bei „Natur_Umwelt: Garten“) oder differenziert (wie z.B. das kulturelle Thema „Kultur: Architektur“ vom technischen Thema „Technik_Industrie: Bau“).

In einigen Fällen konnte jedoch keine einer internen Kategorie entsprechende externe Kategorie gefunden werden, jedenfalls nicht in der erwünschten Feingliedrigkeit; so z.B. bei literarischen Genren. Solche Themen sind in Tabelle 2 mit einem rechtsseitigen Minuszeichen versehen bzw. mit einem Pluszeichen beim gegenteiligen Fall.

Die hier vorgestellte Taxonomie kann als ein Schema zur manuellen Annotierung dienen. Da jedoch – wie eingangs erwähnt – Zeitungstexte wegen ihres Umfangs praktisch nicht manuell annotierbar sind, erfolgte eine automatische Zuweisung, auf die im weiteren Verlauf eingegangen wird. Themen, für die automatisch zugewiesene Zeitungstexte existieren, sind in Tabelle 2 mit einem linksseitigen Pluszeichen markiert bzw. mit einem Minuszeichen bei Eintreten des gegenteiligen Falls.[8]

Hauptkategorie	**Unterkategorie**
Fiktion	*Abenteuer_Historie* (–,–), *Dialekt* (–,–), *Drama* (–,–), *Humor* (–,–), *Krimi* (–,–), *Lyrik_Lieder_Aphorismen* (–,–), *Phantasie_Phantastik* (–,–), *Populärliteratur* (–,–), *Roman_Erzählungen* (–,–), *Sagen_Märchen* (–,–), *Science_Fiction* (–,–), *Tradiertes* (–,–), *Vermischtes* (+,–), *Kinderbücher* (–,–)
Freizeit_Unterhaltung	Fertigkeiten (–,+), Freizeit_Themenparks (–,+), Hausrat (–,+), Reisen (+,+), Rundfunk (+,+), Sonstiges (–,–), Spiele (–,+), Spielsachen (–,+), Täglicher_Bedarf (–,+), Vereine_Veranstaltungen (+,–)
Gesundheit_Ernährung	Ernährung (+,+), Gesundheit (+,+)
Kultur	Architektur (–,+), Bildende_Kunst (+,+), Brauchtum (–,+), Darstellende_Kunst (+,+), Film (+,+), Literatur (+,+), Mode (+,+), Musik (+,+), Sammeln (–,+), Sonstiges (–,–)
Natur_Umwelt	Forst_Agrarwirtschaft (–,+), Garten (+,+), Jagd_Fischfang (–,+), Naturlandschaften (–,+), Sonstiges (–,–), Tiere (+,+), Wetter_Klima (+,+)
Politik	Ausland (+,–), Inland (+,–), Kommunalpolitik (+,–), *Programm* (–,–), *Reden* (–,–), Sonstiges (–,–)
Religion	*Predigt* (–,+), *Quelle* (–,+), Sonstiges (–,–)

[7] Zu einer vollständigen Angabe aller Verankerungen siehe Weiß (2005).

[8] Aus technischen Gründen werden Leerzeichen durch einen Unterstrich wiedergegeben. I.d.R. handelt es sich um sich häufig überschneidende Fassetten von Themen, die formal nicht weiter unterschieden werden, z.B. „Abenteuer_Historie“ für „Abenteuerromane“ und „historische Romane“. In einigen Fällen handelt es sich auch um feste Wendungen, z.B. „Science_Fiction“ für „Science Fiction“.

Sport	Athletik (–,+), Ballsport (+,+), Extremsport (–,+), Fußball (+,+), Kampfsport (–,+), Klettern_Wandern (–,+), Kraftsport (–,+), Motorsport (+,+), Pferde (–,+), Radsport (+,+), Schießen (–,+), Tennis (+,+), Turnen_Gymnastik (–,+), Vermischtes (–,–), Wassersport (–,+), Wintersport (+,+)
Staat_Gesellschaft	Arbeit_und_Beruf (+,+), Bildung (+,+), *Biografien_Interviews* (+,+), Drittes_Reich_Rechtsextremismus (+,+), *Essays* (–,–), Familie_Geschlecht (+,+), Kirche (+,–), Nichtregierungsorganisationen (–,+), Recht (+,+), Sonstiges (–,–), staatliche_Institutionen (–,+), Subkulturen_Minderheiten (–,+), Tod (+,+), Verbrechen (+,+), Vermischtes (–,–)
Technik_Industrie	Bau (–,+), Bergbau_Energie (–,+), Chemie (–,+), Druck_Medien (–,+), EDV_Elektronik (+,+), Handwerke (–,+), Holz (–,+), Keramik_Glas (–,+), Kfz (+,+), Luft_Raumfahrt (–,+), Metall (–,+), Rüstung (–,+), Sonstiges (–,–), Textil_Leder (–,+), Transport_Verkehr (+,+), Umweltschutz (+,+), Unfälle (+,+)
Wirtschaft_Finanzen	Banken (+,+), Bilanzen (+,–), Handel (–,+), Öffentliche_Finanzen (+,–), Sozialprodukt (+,–), Vermischtes (–,–), Währung (+,–)
Wissenschaft	*Musik* (–,+), *Geowissenschaften* (–,+), *Geschichtswissenschaften* (–,+), *Humanwissenschaft* (–,+), *Literaturwissenschaft* (–,+), *Mathematik_Naturwissenschaft* (–,+), *Medizin* (–,+), *Philologie* (–,+), *Populärwissenschaft* (+,–), *Rechtswissenschaft* (–,+), *Sozialwissenschaft* (–,+), *Technik* (–,+), *Theologie_Philosophie* (–,+), *Vermischtes* (–,–), *Wirtschaft* (–,+)
Sonstiges	

Tabelle 2: Thementaxonomie

Kategorie	**externe Definition**

Gesundheit: Ernährung	`http://www.dmoz.org/World/Deutsch/Gesundheit/Ernährung`
	`http://www.dmoz.org/World/Deutsch/Freizeit/Essen_und_Trinken`
	`http://www.dmoz.org/World/Deutsch/Zuhause/Kochen`
	`http://www.dmoz.org/World/Deutsch/Freizeit/Genussmittel`
	`http://www.dmoz.org/World/Deutsch/Wirtschaft/Nahrungs-_und_Genussmittel`

Natur_Umwelt: Garten	`http://www.dmoz.org/World/Deutsch/Zuhause/Garten_und_Pflanzen`
	`http://www.dmoz.org/World/Deutsch/Wirtschaft/Bauwesen/Garten-_und_Landschaftsbau`
	`http://www.dmoz.org/World/Deutsch/Wissenschaft/Naturwissenschaften/Biologie/Botanik/Botanische_Gärten`
	`http://www.dmoz.org/World/Deutsch/Wirtschaft/Konsumgüter/Haus_und_Garten/Garten`

Kultur: Architektur	`http://www.dmoz.org/World/Deutsch/Kultur/Architektur`
Technik_Industrie: Bau	`http://www.dmoz.org/World/Deutsch/Wirtschaft/Bauwesen`

Staat_Gesellschaft: Drittes_ Reich_ Rechtsextremismus	`http://www.dmoz.org/World/Deutsch/Gesellschaft/Politik/Rechtsextremismus`
	`http://www.dmoz.org/World/Deutsch/Gesellschaft/Rassismus_und_Diskriminierung`
	`http://www.dmoz.org/World/Deutsch/Gesellschaft/Geschichte/Nach_Zeitabschitten/Neuzeit/20.Jahrhundert/Nationalsozialismus`
	`http://www.dmoz.org/World/Deutsch/Gesellschaft/Geschichte/Nach_Zeitabschitten/Neuzeit/20.Jahrhundert/Kriege/2.Weltkrieg`

Tabelle 3: Themenkatalog für Sprachkorpora

2.3 Der Klassifikator

Ziel der letzten beiden Abschnitte war die Motivierung einer möglichst umfassenden Thementaxonomie. Dies erfolgte zum einen induktiv, d.h. durch Dokumentclustering, zum anderen deduktiv in Form einer Anpassung einer bereits bestehenden Taxonomie. Das somit spezifizierte Klassifikationsschema ist gegenüber den in der Einleitung erwähnten Schemata objektiver, da es in einer externen Ontologie verankert ist. Es ist zudem auch intuitiver, da es datenbasiert ist, insofern als für die meisten Kategorien Indikatoren in Form von Clustern geliefert werden.

Ziel dieses Abschnittes ist, neben einer manuellen Klassifikation auch eine vollautomatische zu ermöglichen. Dies geschieht unter Zuhilfenahme eines Textklassifikators. Textklassifikation kann als die Wahl einer Klasse für ein a priori unklassifiziertes Dokument unter Berücksichtigung von a priori klassifizierten Dokumenten bzw. Trainingsdaten beschrieben werden. Als Trainingsdaten dienen die im vorletzten Abschnitt beschriebenen Cluster.

Da hier deren Umfang sehr groß ist und die Daten wegen des semiautomatischen Konstruktionsverfahrens keine hundertprozentige Korrektheit aufweisen, wurde ein möglichst robustes Klassifikationsverfahren gewählt. Als ein solches gilt die *Naive Bayes'sche Textklassifikation*. Das Verfahren, das z.B. in Mitchell (1997) beschrieben wurde, ist probabilistisch, insofern als die Wahrscheinlichkeit eines Dokumentes v zu einem Thema θ sich als die multiplikative Verknüpfung der (aus den Trainingsdaten entnommenen) Wahrscheinlichkeiten der Terme in v zu θ errechnen lässt.[9] Im Gegensatz zum Clusterer wurde für den Klassifizierer kein Lemmatisierer verwendet, da auch morphologische Aspekte Rückschlüsse auf Kategorien ermöglichen, z.B. in Formen der Vergangenheitsangabe wie „Präteritum“ oder „Perfekt“. Darüber hinaus wurde auch auf eine Stoppwortliste verzichtet, da Präferenzen für an und für sich unbedeutende Wörter, z.B. für das Pronomen „ich“ gerade bei schwer zu klassifizierenden Themen wie „Belletristik“, „Lebensläufen“ oder „wissenschaftlichen Texten“ wichtig sind. Es wurden sogar Formatierungsangaben wie Absätze mitberücksichtigt, da sich zeigte, dass gerade solche Angaben bei schwierigen Entscheidungen wie z.B. der zwischen einer „Kriminalgeschichte“ („Fiktion: Vermischtes“) und einer Kriminalchronik („Staat_Gesellschaft: Verbrechen“) für die richtige Vorhersage sorgen.

Aus verarbeitungstechnischen Gründen wurden nur Terme mit mehr als fünfmaligem Vorkommen zugelassen. Im Gegensatz zum klassischen Verfahren wurde auf das Einbeziehen von Häufigkeitsverteilungen der Dokumente verzichtet. Der Grund hierfür liegt in der offenen thematischen Zusammensetzung des Korpus nach weiteren Akquisitionen.

3. Die Auswertung

Für die Evaluierung der Daten wurden *Präzision*, d.h. der relative Anteil der korrekten Cluster- und Klassifikationsergebnisse in Bezug auf das jeweilige Gesamtergebnis und *Recall*, d.h. der relative Anteil korrekter Klassifikationsergebnisse in Bezug auf eine extern klassifizierte Datenmenge berechnet. Diese Datenmengen wurden durch Zufallsauswahl extrahiert und umfassten je 30 Dokumente pro Kategorie. *Präzision* ist in Tabelle 4 dargestellt. Die Berechnung der Präzision erfolgte zunächst separat für jedes Unter- und Oberthema. Hieraus wurde dann die durchschnittliche Präzision für Unter- und Oberklassen errechnet. Ersteres fungiert als allgemeine Präzisionsangabe. Wie aus der Tabelle ersichtlich, wurde dieses Be-

[9] Dieses Verfahren gilt als „naiv“, da es von *statistischer Unabhängigkeit* ausgeht; d.h., die Wahrscheinlichkeit des Auftreten eines Terms gilt als unabhängig von der Auftretenswahrscheinlichkeit eines anderen Terms. Dies würde bedeuten, dass beispielsweise die Wahrscheinlichkeit des Auftretens von „Bundeskanzler“ unabhängig ist vom Auftreten thematisch verwandter Terme wie z.B. „Bundesregierung“, was klar nicht der Fall ist. Trotzdem gilt die *Naive Bayes'sche Textklassifikation* als eine der erfolgreichsten (siehe Domingos/ Pazzani 1996).

rechnungsverfahren jeweils für den Clusterer (siehe Spalte TRK-B) und den Klassifikator (Spalte TRK-K, NTR-K, 03-K) durchgeführt. Für letzteres wurde unterschieden zwischen Trainingsdaten (Spalte TRK-K), Daten, die keine Trainingsdaten darstellen, aber zu Jahrgängen gehören, für die Trainingsdaten vorhanden sind (Spalte NTR-K) und schließlich „absoluten Neudaten“, d.h. Daten, die weder Trainingsdaten darstellen, noch zu Jahrgängen gehören, für die Trainingsdaten existieren. Dies betrifft Daten des Jahrgangs 2003, daher die Bezeichnung 03-K. Bezüglich der Interpretation von Tabelle 4 gibt es drei Auffälligkeiten: Erstens kann ein Gefälle zwischen Trainings- und Nichttrainingsdaten beobachtet werden. Eine Ausnahme stellt die Kategorie „Sport: Motorsport“ dar, was jedoch durch den hohen Anteil an „Schrottdokumenten“ im Trainingskorpus erklärt werden kann. Die zweite, eher überraschende Beobachtung betrifft die Tatsache, dass die Spalte TRK-K (das Ergebnis des Klassifikators für die Trainingsdaten also) Spalte TRK-B (d.h. das Ergebnis des Clusterers) übertrifft, jedenfalls im Durchschnitt. Diese Beobachtung deckt sich jedoch mit anderen Anwendungen der naiven Bayes'schen Textklassifikation, wie z.B. beschrieben in Lang (1995). Die dritte Beobachtung, nämlich dass 03-K wider Erwarten bessere Ergebnisse liefert als NTR-K, kann aus der Tatsache erklärt werden, dass sich diese Korpora aus unterschiedlichen Zeitungen zusammensetzen. So kann beispielsweise das wesentlich bessere Abschneiden von „Sport: Wintersport“ damit erklärt werden, dass Zeitungen, die über mehrere Sportarten, wie z.B. „Skispringen“ und „Handball“, gleichzeitig berichten, vermehrt in NTR-K auftreten und in 03-K fehlen. Die Korpora NTR-K und 03-K sind also nicht als ganz äquivalent anzusehen.

Bezüglich der Berechnung des Recalls sei auf Tabelle 5 verwiesen. Es wurde wie folgt vorgegangen: Durch Zufallsauswahl und anschließende manuelle Kontrolle wurden für jede Kategorie 30 korrekte Texte aus dem Trainingskorpus extrahiert. Der relative Anteil der jeweiligen Klassifikationsergebnisse in Bezug auf diese Extraktion bildete das Ergebnis. Es konnte beobachtet werden, dass Fälle von geringem Recall sich auf Themen beschränkten, für die zum einen nur wenige Trainingsdaten gefunden wurden und/oder für die es ähnliche Kategorien gab, die jedoch mehr Daten umfassten. Letzteres konnte an Hand einer „Konfusionsmatrix“ nachgewiesen werden. Es ist daher damit zu rechnen, dass die Akquisition von zusätzlichen Daten bessere Ergebnisse liefert.

Thema	**TRK-B**	**TRK-K**	**NTR-K**	**03-K**
Fiktion: Vermischtes	1,000000	1,000000	0,700000	0,812500
Freizeit_Unterhaltung: Reisen	0,933333	0,933333	0,533333	0,677419
Freizeit_Unterhaltung: Rundfunk	1,000000	1,000000	0,900000	0,968750
Freizeit_Unterhaltung: Vereine_Veranstaltungen	0,733333	1,000000	0,741935	0,838710
Gesundheit_Ernaehrung: Ernaehrung	0,933333	1,000000	0,833333	0,962963
Gesundheit_Ernaehrung: Gesundheit	0,966667	1,000000	0,866667	0,966667
Kultur: Bildende_Kunst	0,966667	0,967742	0,833333	0,967742
Kultur: Darstellende_Kunst	1,000000	0,966667	0,800000	0,750000
Kultur: Film	1,000000	0,966667	0,733333	0,852941
Kultur: Literatur	0,900000	0,866667	0,666667	0,700000
Kultur: Mode	0,900000	1,000000	0,866667	0,973684
Kultur: Musik	0,966667	0,966667	0,766667	0,838710
Natur_Umwelt: Garten	1,000000	1,000000	0,666667	0,657143
Natur_Umwelt: Tiere	0,966667	1,000000	1,000000	0,941176
Natur_Umwelt: Wetter_Klima	0,933333	1,000000	0,933333	1,000000
Politik: Ausland	0,933333	1,000000	0,866667	0,935484
Politik: Inland	0,833333	0,969697	0,866667	0,878788
Politik: Kommunalpolitik	0,966667	1,000000	0,968750	0,933333
Sport: Ballsport	0,733333	0,933333	0,866667	0,533333
Sport: Fussball	0,900000	1,000000	0,833333	0,677419
Sport: Motorsport	0,766667	0,800000	0,933333	0,933333
Sport: Radsport	0,966667	0,933333	0,900000	0,900000

Sport: Tennis	1,000000	0,966667	0,833333	0,800000
Sport: Vermischtes	0,733333	0,833333	0,766667	0,766667
Sport: Wintersport	0,866667	0,800000	0,566667	0,833333
Staat_Gesellschaft: Arbeit_und_Beruf	0,933333	1,000000	1,000000	1,000000
Staat_Gesellschaft: Bildung	0,966667	1,000000	0,966667	0,900000
Staat_Gesellschaft: Biographien_Interviews	0,933333	0,709677	0,620690	0,612903
Staat_Gesellschaft: Drittes_Reich_Rechtsextremismus	0,966667	1,000000	0,633333	0,677419
Staat_Gesellschaft: Familie_Geschlecht	0,833333	0,800000	0,700000	0,580645
Staat_Gesellschaft: Kirche	1,000000	0,968750	0,666667	0,866667
Staat_Gesellschaft: Recht	1,000000	1,000000	0,966667	1,000000
Staat_Gesellschaft: Verbrechen	0,733333	0,933333	0,866667	0,966667
Technik_Industrie: EDV_Elektronik	1,000000	0,900000	0,866667	0,533333
Technik_Industrie: Kfz	0,866667	0,900000	0,900000	1,000000
Technik_Industrie: Transport_Verkehr	0,966667	0,906250	0,866667	0,833333
Technik_Industrie: Umweltschutz	0,966667	0,966667	0,866667	0,766667
Technik_Industrie: Unfaelle	0,866667	1,000000	0,966667	0,900000
Wirtschaft_Finanzen: Banken	0,900000	1,000000	0,666667	0,466667
Wirtschaft_Finanzen: Bilanzen	0,900000	0,967742	0,933333	0,933333
Wirtschaft_Finanzen: Oeffentliche_Finanzen	0,866667	0,933333	0,700000	0,933333
Wirtschaft_Finanzen: Sozialprodukt	0,700000	0,870968	0,866667	0,866667
Wirtschaft_Finanzen: Waehrung	0,666667	0,900000	0,733333	0,812500
Wissenschaft: Populaerwissenschaft	0,833333	0,966667	0,620690	0,875000
Fiktion	1,000000	1,000000	0,700000	0,812500
Freizeit_Unterhaltung	0,888889	0,977778	0,725275	0,840426
Gesundheit_Ernaehrung	0,950000	1,000000	0,850000	0,964912
Kultur	0,955556	0,966851	0,844444	0,908163
Natur_Umwelt	0,966667	1,000000	0,866667	0,876190
Politik	0,911111	1,000000	0,934783	0,936170
Sport	0,895238	0,928571	0,933333	0,862559
Staat_Gesellschaft	0,929630	0,945255	0,840149	0,870445
Technik_Industrie	0,933333	0,947712	0,926667	0,866667
Wirtschaft_Finanzen	0,933333	0,974359	0,906667	0,921053
Wissenschaft	0,833333	0,966667	0,620690	0,875000
Klassen	**0,906667**	**0,946942**	**0,814952**	**0,832847**
Oberklassen	0,927407	0,962417	0,865285	0,885401
Schrott	0,876190	0,995261	0,952153	0,988827
Anteil_an_falschen_Schrottdokumenten	0,023704	0,016212	0,020725	0,027007

Tabelle 4: Präzisionstabelle für klassifiziertes Korpus

Thema	Korrekt (prozentualer Anteil)
Fiktion: Vermischtes	63,33
Freizeit_Unterhaltung: Reisen	80,00
Freizeit_Unterhaltung: Rundfunk	63,33
Freizeit_Unterhaltung: Vereine_Veranstaltungen	80,00
Gesundheit_Ernaehrung: Ernaehrung	0,00
Gesundheit_Ernaehrung: Gesundheit	90,00
Kultur: Bildende_Kunst	100,00
Kultur: Darstellende_Kunst	86,67
Kultur: Film	80,00
Kultur: Literatur	90,00
Kultur: Mode	46,67
Kultur: Musik	96,67
Natur_Umwelt: Garten	13,33
Natur_Umwelt: Tiere	63,33
Natur_Umwelt: Wetter_Klima	90,00
Politik: Ausland	93,33
Politik: Inland	93,33
Politik: Kommunalpolitik	90,00
Schrott: boersenkurse	93,55
Schrott: geburt_tod_heirat	90,00
Schrott: impressum	90,00
Schrott: inhaltsverzeichnisse	80,00
Schrott: ligatabellen	96,67
Schrott: tabellen	93,33
Schrott: veranstaltungshinweise	86,67
Sport: Ballsport	60,00
Sport: Fussball	100,00
Sport: Motorsport	100,00
Sport: Radsport	96,67
Sport: Tennis	100,00
Sport: Vermischtes	83,33
Sport: Wintersport	96,67
Staat_Gesellschaft: Arbeit_und_Beruf	93,33
Staat_Gesellschaft: Bildung	93,33
Staat_Gesellschaft: Biographien_Interviews	93,33
Staat_Gesellschaft: Drittes_Reich_Rechtsextremismus	70,00
Staat_Gesellschaft: Familie_Geschlecht	80,00
Staat_Gesellschaft: Kirche	83,33
Staat_Gesellschaft: Recht	90,00
Staat_Gesellschaft: Tod	100,00
Staat_Gesellschaft: Verbrechen	90,00
Technik_Industrie: EDV_Elektronik	90,00
Technik_Industrie: Kfz	96,67
Technik_Industrie: Transport_Verkehr	76,67
Technik_Industrie: Umweltschutz	90,00
Technik_Industrie: Unfaelle	100,00
Wirtschaft_Finanzen: Banken	86,67
Wirtschaft_Finanzen: Bilanzen	90,00
Wirtschaft_Finanzen: Oeffentliche_Finanzen	76,67
Wirtschaft_Finanzen: Sozialprodukt	80,00
Wirtschaft_Finanzen: Waehrung	80,00
Wissenschaft: Populaerwissenschaft	80,65
Durchschnitt (Schrott)	**90,03**
Durchschnitt (ohne Schrott)	**82,16**

Tabelle 5: Angabe des Recalls für das Trainingskorpus

4. Sonstige Ergebnisse

Neben dem Hauptziel, der thematischen Annotierung von Dokumenten eines Sprachkorpus, möchte ich in diesem Kapitel kurz weitere ergänzende Punkte ansprechen.[10] Diese sind:

4.1 Schlüsselwortextraktion

Unter einem *Schlüsselwort* verstehe ich einen Term, der in Bezug auf ein Thema signifikant häufig auftaucht. Diese Signifikanz kann durch den χ^2-Test formal bestimmt und graduiert werden. Tabelle 6 zeigt für eine Auswahl von Themen die jeweils zehn signifikantesten Terme. Wie aus der Tabelle hervorgeht, besteht die Funktion des Schlüsselwortes darin, Rückschlüsse bezüglich der Inhalte von Textmengen „auf einen Blick“ zu ermöglichen.

Thema	**Schlüsselwörter**
Kultur: Darstellende_Kunst:	theaters (140861,327926), theater (106456,298078), inszenierung (42258,945867), schauspieler (41939,399374), intendant (40583,800429), buehnen (39087,910553), buehne (36909,269587), inszenierungen (35637,136278), intendanten (28001,356477), regisseur (26609,299917)
Politik: Inland:	spd (88656,475356), cdu (69442,202956), partei (60557,279373), koalition (55030,591700), fdp (47300,666759), gruenen (43418,271046), pds (35011,049061), bundeskanzler (30870,169622), kanzler (30180,140795), sozialdemokraten (29041,860785)
Staat_Gesellschaft: Arbeit_und_Beruf:	ig (112078,730976), metall (97948,439509), gewerkschaft (90792,839414), arbeitgeber (81908,649540), gewerkschaften (54783,865044), arbeitgebern (45906,969656), streik (43353,549910), arbeitskampf (42472,160896), beschaeftigten (41169,045675), tarifverhandlungen (40947,802934)
Staat_Gesellschaft: Drittes _Reich_Rechtsextremismus:	neonazis (76584,092859), nazis (45013,193609), neonazi (34488,591613), juden (19401,930548), rechtsextremisten (14575,331525), nazi (13452,855226), juedischen (12954,927297), nazi-opfer (11304,944080), nationalsozialisten (11008,634908), worch (10031,533364)
Staat_Gesellschaft: Verbrechen:	taeter (109429,787550), polizei (66255,586435), einbrecher (48550,594389), beamten (42406,380701), raeuber (41651,657443), beute (38832,242147), bargeld (28927,563005), unbekannte (28499,462353), ueberfall (28434,994193), ueberfallen (24369,433529)
Technik_Industrie: EDV_Elektronik:	windows (105095,338706), microsoft (101588,363129), software (75859,723123), betriebssystem (65576,725654), internet (63063,471347), anwender (54070,460275), pc (43526,762328), server (39983,309524), online (39199,589238), nt (36930,037004)

Tabelle 6: Durch den χ^2-Test gewonnene Schlüsselwörter

4.2 Verwendung von Clustern zur Recherchezwecken

Wie in Abschnitt 2.1 beschrieben, erfolgte der Aufbau einer Thementaxonomie bzw. eines Trainingskorpus über den Einsatz eines Clusterverfahrens. Neben dieser mittelbaren Funktion können Cluster jedoch auch eine unmittelbare erfüllen, nämlich Texte zu konkreten Ereignissen zu liefern, z.B. „die Wende(n) in Ost- und Mitteleuropa um das Jahr 1990“ oder „EU-Erweiterungsverhandlungen“. Ein Ausschnitt von solchen Ereignissen kann der linken Spalte in Tabelle 1 entnommen werden.

[10] Aus Platzgründen werde ich nur Ausschnitte der jeweiligen Ergebnisse vorstellen. Für eine vollständige Übersicht verweise ich auf Weiß (2005).

4.3 Filterung von Datenmüll

Zeitungen, aus denen Korpora zu einem Großteil bestehen, enthalten sehr viel linguistisch Uninteressantes wie *Ligatabellen, Börsenkurse, Veranstaltungshinweise* usw. Um das Erscheinen dieser unerwünschten Dokumente zu unterbinden, wurden analog zum beschriebenen Verfahren entsprechende Trainingsdaten spezifiziert. Dokumente, die zu dieser Klasse gehören, werden unterbunden. Bezüglich der Evaluierung wurde neben der Präzision auch der Anteil an Dokumenten berechnet, die fälschlicherweise nicht als Datenmüll kategorisiert wurden. Dieser Anteil ist in der untersten Zeile von Tabelle 4 vermerkt.

5. Zusammenfassung

Ziel dieses Papiers war die Beschreibung des Teilprojekts „Thematische Erschließung". Ich ging zunächst auf bestehende Kategorisierungsschemata von Sprachkorpora ein und konstatierte folgende Mängel:

- Einige bestehende Ansätze sind fast ausschließlich an Textgattungen orientiert und nicht an Inhalten. Bei Ansätzen, die mehr an Inhalten orientiert sind, fehlt es meiner Meinung nach an Ausgewogenheit.
- Die Zuweisung von Kategorien funktioniert rein manuell, was die thematische Annotierung von Massendaten wie z.B. Zeitungstexten ausschließt.

Mein Vorschlag, diese Mängel zu beheben, erfolgt in einem zweistufigen Modell: In der ersten Stufe wird in einem semiautomatischen Verfahren unter Zuhilfenahme eines Dokumentclusterers und einer externen, höherstufigen Ontologie eine Taxonomie von Themen spezifiziert. Dieses Verfahren ist sowohl „intuitiver", da es datenbasiert ist, als auch „objektiver" insofern es in einer externen Ontologie verankert ist. Die zweite Stufe besteht im Einsatz eines Textklassifikators, durch den es möglich wird, jeden Text eines Korpus zu klassifizieren. Beide Verfahrenstufen wurden formal evaluiert und weitere Anwendungen angesprochen.

Literatur

Belica, Cyril (1994): *MLAP93-21 MECOLB*. Deliverable D5. WP2 – Lemmatizer. Final Report, July 1994. Task 2.1: Lemmatizer. Luxembourg: MECOLB. Internet: `www.ids-mannheim.de/kt/dokumente/glemmrep.pdf` (Stand: Dezember 2005).

Domingos, Pedro/Pazzani, Michael (1996): Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. In: Saitta, Lorenza (Hg.): Machine Learning: Proceedings of the Thirteenth International Conference (ICML '96), Bari, Italy, July 3-6, 1999. San Francisco, CA: Morgan Kaufmann. S. 105-112.

Jain, Anil K./Dubes, Richard C. (1988): Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall.

Karypis, George (2002): CLUTO A Clustering Toolkit. Technical Report 02-017. Minneapolis, MN: Dept. of Computer Science, University of Minnesota. Internet: `www.cs.umn.edu/~cluto` (Stand: Dezember 2005).

Karypis, George/Zhao, Ying (2002): Comparison of Agglomerative and Partitional Document Clustering Algorithms. Manual. Minneapolis, MN: Dept. of Computer Science, University of Minnesota. Internet: `www.ece.northwestern.edu/~yingliu/datamining_papers/paper2.pdf` (Stand: Dezember 2005).

Kučera, Henry/Francis, W. Nelson (1979): Brown Corpus Manual. Revised and Amplified Version 1979. Providence, RI: Department of Linguistics, Brown University. Internet: `http://nora.hd.uib.no/icame/brown/bcm.html` (Stand: Dezember 2005).

Lang, Ken (1995): Newsweeder: Learning to Filter Netnews. In Prieditis, Armand/Russell, Stuart (Hg.): Proceedings of the Twelfth International Conference on Machine Learning: June 12-14, 1995, Tahoe City, California. San Francisco, CA: Morgan Kaufmann. S. 331-339.

Mitchell, Tom. M. (1997): Machine Learning. Boston, MA: McGraw-Hill.

Netscape (2002): About the Open Directory Project. Internet: www.dmoz.org/about.html (Stand: Dezember 2005).

PAROLE (2004): Information on the PAROLE Corpus. Internet: http://parole.inl.nl/html-eng/main_info.html (Stand: Dezember 2005).

Salton, Gerard (1968): Automatic Information Organization and Retrieval. New York/London: McGraw-Hill.

Spark-Jones, Karen (1972): A Statistical Interpretation of Term Specifity and its Application in Retrieval. In: Journal of Documentation 28, S. 11-21.

Weiß, Christian (2005): Dokumentation zu dem Projekt: Thematische Erschließung. Internet: www.ids-mannheim.de/kt/projekte/methoden/te.html (Stand: Dezember 2005).