In: Elisabeth Niggemann, Klaus Lepsky (Hrsg., 1996): Zukunft der Sacherschließung im OPAC. Vorträge des 2. Düsseldorfer OPAC-Kolloquiums. Schriften der Universitäts- und Landesbibliothek Düsseldorf Band 25, S. 37-48

Automatische Indexierung und elektronische Thesauri

Harald H. Zimmermann

1. Einführung

Die Diskussion über den praktischen Wert des Einsatzes sprachanalytischer Verfahren in der Erschließung und Verfügbarmachung von Texten (und Titeln) ist fast so alt wie die Erstellung elektronischer Textdatenbanken bzw. bibliographischer Datenbanken (Literaturdatenbanken).
Gerald Salton, zunächst ein Pionier des sprachanalytischen Information Retrieval, hatte sich aufgrund wenig überzeugender Ergebnisse zunehmend auf rein statistische Konzepte (v.a. des Document Clustering) verlegt. Vor allem bei Datensammlungen, denen englischsprachige Texte zugrunde liegen, schien die simple Vorgehensweise einer Rechtstrunkierung beim Retrieval, verbunden mit der Ausnutzung von Wortabständen (Adjacency- und Same-Funktion) angesichts der an sich schon großen Variationsbreite sprachlicher Ausdrucksformen, die von "höherwertigen" Systemen auch nicht bewältigt wurden, ein angemessenes Verfahren.

In Deutschland gab es allerdings - bedingt durch die stärkere Flexion deutschsprachiger Wörter, v.a. auch durch die Problematik der Wortzusammensetzungen - schon frühzeitig die Bemühung, Verfahren zu entwickeln, die zumindest an der Oberfläche eine grundformorientierte Indexierung ermöglichten und darüber hinaus bei den Komposita eine Dekomposition in (möglichst sinnvolle) Wortbestandteile erzielten. Das von Siemens entwickelte System PASSAT kann hier als Beleg dienen.

Während PASSAT "einzelwortorientiert" operierte, wurde mit dem an der Universität des Saarlandes entwickelten System CTX (Computergestützte Texterschließung) ein syntaxbasierter Ansatz verfolgt: Neben einer Grundformenermittlung, ebenfalls verbunden mit einer Dekomposition, wurde über eine syntaktische Analyse versucht, Mehrwortgruppen zu erschließen, wobei alleine die Strukturanalyse (etwa die Verbindung "Adjektiv+Substantiv" bzw. "Substantiv+Präposition+Substantiv") für die Erschließung einer solchen "Gruppe" herangezogen wurde.

In einem groß angelegten Projekt des BMFT wurde anhand von Daten aus dem Patentbereich ein erster Vergleich der verschiedenen Verfahren (PASSAT, CTX, "reine" wortformenbasierte Erschließung) durchgeführt, wobei sowohl ein "spezifisches" Recall/Precision-Maß benutzt als auch das Handling und die Aufwendungen in Betracht gezogen wurden. Die beiden linguistischen Verfahren schnitten im Gesamtergebnis deutlich besser ab, wobei PASSAT v.a. in Bezug auf das Handling und den erforderlichen Aufwand noch Vorteile bot.

Damit war im Grundsatz der Nachweis gelungen, dass sich zumindest beim Deutschen sprachanalytische Verfahren lohnen. Es war zugleich deutlich geworden, dass sich weitergehende, syntaxbasierte Systeme noch nicht "rechneten" (geschweige denn Ansätze, die das - an sich sehr wichtige - Problem der automatischen semantischen Disambiguierung zu lösen vermöchten).

An dieser Stelle soll zumindest auf eine weitere Möglichkeit linguistisch basierter Erschließungen hingewiesen werden, die sich ebenfalls heute in der Praxis schon bewährt hat: In der Literaturdokumentation, bei der sprachliche Inhaltsangaben (sog. Abstracts) als (maschinenlesbare) Quellen für weiterverarbeitende Verfahren zugrunde gelegt werden können, hat das ursprünglich von G. Lustig in Darmstadt entwickelte Indexierungssystem gezeigt, dass es möglich ist, die intellektuelle (verdichtende) Indexierung maschinell zu unterstützen, indem systemseitig Vorschläge für die (thesaurusgebundene) Vergabe von Deskriptoren gemacht werden. Ein solches Verfahren ist allerdings für OPACs, denen meist nur kurze Titeltexte zugrunde liegen, mangels Volumen nicht machbar.[1]

2. Das IDX-Verfahren

Im Rahmen der Vorstudien zu MILOS war erkennbar geworden, dass das von der Firma SOFTEX entwickelte Verfahren "IDX" eine gute Grundlage für die Erschließung von Titeln für den Düsseldorfer OPAC darstellt.[2] Die für die folgende Darstellung wesentlichen Komponenten sollen hier nur kurz zusammengefasst werden (es wird nur das Verfahren für die deutsche Sprache behandelt, obgleich IDX in analoger Form auch für die Sprachen Französisch und Englisch sowie inzwischen für Spanisch und Italienisch verfügbar ist):

(1) Textwortformen (z.B. Kindes, Häusern, gegangen ...) werden automatisch auf die relevante "Grundform" (hier: Kind, Haus, gehen) zurückgeführt. Grundlage ist ein Deflexionsverfahren, das im wesentlichen entsprechende Markierungen der Wortstämme des sog. "Rechtschreibwörterbuchs" auswertet.

Um "voll" wirksam zu sein, muss dieses Wörterbuch bei einem Textabgleich entsprechend erweitert werden, soweit neue Wörter auftreten, die entsprechender Markierungen bedürfen. Für den "Allgemeingebrauch" sind die bestehenden Einträge hinreichend (Standardumfang rd. 120.000 "nutzbare" Stämme), für spezifische Anwendungen (wie in der heterogenen OPAC-Welt der ULB Düsseldorf) sind allerdings Erweiterungen erforderlich.

(2) Bei Wortzusammensetzungen und -ableitungen (z.B. Haustüren, Kinderzimmer, Besichtigungen ...) werden die Grundform sowie die Teilwörter ermittelt (Haustür => Tür; Haustür => Haus; Kinderzimmer => Kind, Kinderzimmer => Zimmer; Besichtigung => besichtigen).

Grundlage sind einerseits Merkmale im "Rechtschreibwörterbuch" (zur Dekomposition), die über eine Wortanalyse automatisch aufgrund lexikalischer Bestandteile ermittelt werden, andererseits Einträge in einem sog. "Relationenwörterbuch", die u.a. auch formal mögliche, sprachlich "sinnlose" Teilwortrelationen zu differenzieren erlauben.

Da ein Eintrag im "Relationenwörterbuch" Priorität vor einer automatischen Zerlegung hat, kann über eine entsprechende Pflege des Relationenwörterbuchs sichergestellt werden, dass fehlerhafte Lösungen (später) vermieden werden. Das derzeitig systemseitig verfügbare Relationenwörterbuch hat ein Volumen von rd. 120.000 Relationen vorwiegend vom Typ Kompositum => Teilwort (Invertierungen nicht gerechnet).

(3) Soweit die Einträge (auf Grundformbasis) lexikalisiert sind, können auch (im Text kontinuierlich auftretende) sog. "Mehrwortbegriffe" bei der Textanalyse als solche identifiziert werden.[3] Lexikalische Grundlage können dabei das schon erwähnte Relationenwörterbuch oder aber ein entsprechend strukturiertes "Übersetzungswörterbuch" sein. Typische "lexikalisierte" Mehrwortbegriffe sind Wörter wie "juristische Person", "Schutz personenbezogener Daten" usf. Sie lassen sich ebenfalls mit ihren (unmittelbaren) Teilwörtern in Beziehung setzen.

Bei dieser Gelegenheit kann eine weitere Spezifität der Relationierung kurz erläutert werden, die allgemein "mehrstufige Relationierung" heißt: Ein Eintrag wie "politische Willensbildung" wird auf der "ersten" (obersten) Stufe "nur" mit den Teilwörtern "politisch" und "Willensbildung" in Beziehung gesetzt, diese wiederum werden auf einer "2. Stufe" (aus der "Sicht des o.a. Mehrwortbegriffs) weiter relationiert, etwa "politisch" zu "Politik" (Derivation) und "Willensbildung" zu "Wille" bzw. "Bildung" (zur Problematik der Bedeutungsunterscheidung s.u.). Es bleibt allerdings dem Nutzer selbst überlassen, diese "Stufen" zu bestimmen (es kommt also ganz darauf an, wie ein System "gepflegt" oder adaptiert wird).

Bei dem Indexierungsvorgang (Titelanalyse) wird ggf. - je nach Bedarf des Nutzers - die jeweilige "Indexierungstiefe" über Parameter aktiviert oder deaktiviert.

(4) Bis zu einem gewissen Identifikationsgrad (hier gibt es also keine 100%-Lösung) können derzeit zwei weitere - auf der Textanalyse aufbauende - Verfahren genutzt werden:

(a) Auflösung von Teilwort-Tilgungen: Hierbei handelt es sich um Fälle wie "Haus- und Hofwirtschaft", „Kinder- oder Jugendbücher", bei denen einerseits die ganze "Gruppe" als solche identifiziert wird, ohne "lexikalisiert" zu sein, andererseits auch versucht wird, das "getilgte" Teilwort zu erkennen (in den obigen Beispielen sind dies "Hauswirtschaft" bzw. "Kinderbuch").

(b) Soweit sie "lexikalisch" zugelassen sind, können diskontinuierliche Verbteile dem Hauptbestandteil zugeordnet werden (Beispiel: kamen ... an => ankommen).

Diese Verfahren sind derzeit im Ausbau begriffen. Insbesondere soll versucht werden, aufgrund besonderer Kontextwörter (etwa Berufsbezeichnungen, Vornamenlisten) Eigennamen und Vor- und Familiennamengruppen auch dann als solche zu identifizieren, wenn sie (noch) nicht lexikalisiert sind. Analog wird versucht, aus dem Kontext heraus Produkt- und Verfahrensnamen zu erschließen.

Theoretisch ist es aufgrund des zugrunde liegenden partiellen Parsing-Verfahrens auch möglich, nicht-lexikalisierte Mehrwortgruppen (etwa der Form "Adjektiv + Substantiv" oder "Substantiv + Präposition + Substantiv" ...) zu identifizieren. Diese Methode wird inzwischen im Vorfeld einer Indexierung für den Aufbau möglicher Mehrwortgruppen bereitgestellt: Große Datenmengen werden zunächst nach "potentiellen Kandidaten" für Mehrwortgruppen durchforstet, die dann nach "statistischer" Filterung zur Kontrolle bereitgestellt werden, wobei entschieden werden kann, ob es sich nur um "zufällige" Häufigkeiten oder aber um sinnvolle Mehrwortgruppen handelt.

3. Einige grundlegende Betrachtungen zum Thesaurusbegriff

In der "klassischen" Form (zumindest, soweit man die DIN-Schriften zugrunde legt) hat ein Thesaurus folgende Merkmale:

Einerseits handelt es sich um eine kontrollierte, d.h. auf Zulässigkeit "gesichtete" und ggf. durch Definitionen oder erläuternde Zusätze disambiguierte Liste von Wörtern einer natürlichen Sprache.

Die Liste dient dem Zweck der inhaltlichen Beschreibung eines Textes bzw. Dokuments, der sog. „intellektuellen" Indexierung (nicht zu verwechseln mit der automatischen Voll- oder Freitextindexierung). Bei dieser Art von Indexierung wird der Inhalt des Dokuments durch die Auswahl der zugelassenen Wörter aus der Liste repräsentiert. Anders gesagt: der Inhalt eines ganzen Buches oder Aufsatzes wird - formal gesehen - auf die begrifflichen Inhalte abgebildet, die diesen Wörtern zugeordnet sind.

Ein weiteres Merkmal des Thesaurus ist, dass die Wörter, genauer: die Terme (da dahinter aufgrund der Disambiguierung jeweils nur noch eine "Bedeutung" steht) zueinander in (verschiedene) Beziehung(en) gesetzt sind. Diese "Relationen" können sehr differenziert sein; ein "klassischer" Thesaurus verzeichnet zumindest Synonymbeziehungen und Oberbegriff/Unterbegriff-Beziehungen zwischen Inhalten (die mit den Wörtern verbunden sind). Eine besondere Art der Relationierung (die v.a. aus der Problematik entstanden sein dürfte, dass gedruckte Thesauri ansonsten sehr umfangreich geworden und zudem auch die Indexierungsergebnisse nur schwer nutzbar gewesen wären) ist die Beziehung zwischen einem sog. "Deskriptor" (als einem zur Indexierung allein erlaubtem Term/Wort) und einem "Nichtdeskriptor" (als nicht erlaubtem Wort, an dessen Stelle der relationierte Term zu verwenden ist).

Man muss damit einen Thesaurus einerseits als Instrument eines Indexierungssystems verstehen (Ziel ist die Verdichtung häufig komplexer Sachverhalte auf meist nur wenige "Kernthemen"), andererseits ist sein Geltungsbereich in der Regel auf einen mehr oder weniger umfangreichen Wissensbereich begrenzt.

Ein Thesauruselement (Deskriptor) steht - vergleichbar mit einem Element einer nicht-natürlichsprachigen Notation (kurz "Klassifikation") - nicht für die Bedeutung des Wortes "an sich", sondern für einen Themenbereich. Damit hat der Thesaurus mit einem Klassifikationssystem eine Problematik gemeinsam: die der Feingliederung, d.h. der Differenzierung der Sachverhalte (Themen) selbst.

Man kann sich dies sehr einfach an einem Beispiel klarmachen: Angenommen, das Themengebiet des Thesaurus sei die Religionswissenschaft, der praktische Bearbeitungsschwerpunkt (bedingt durch die der Erschließung zugrundeliegenden Materialien) sei die christliche Religion. Es wird weiter angenommen, dass für die Bearbeitung dieses Bereiches ausreichend Experten zur Verfügung stehen. Die Nutzer sind ebenfalls Experten im engeren Fachgebiet, die Materialmenge verlangt eine weitgehende Differenzierung. Man wird einen Thesaurus (oder eine Klassifikation) hier sehr weit auffächern, bei der Deskribierung eines Randgebietes aber allenfalls "allgemeine"

Wörter verwenden (etwa einerseits "evangelische Katechetik", andererseits "Buddhismus"). Umgekehrt würde ein Thesaurus, der Datenmaterial an einer indischen Hochschule beschreibt, den Themenbereich "Buddhismus" für seine Klientel weiter auffächern, bei der Beschreibung christlicher Literatur andererseits mit einem "allgemeinen" Wort auskommen.

Ein Hauptargument für die Anwendung einer Indexierung mit Hilfe eines Thesaurus - dies ist für die folgende Argumentation besonders wichtig - ist die "leichtere" Verwendbarkeit beim Retrieval. Hierunter versteht man die Informationssuche in einer Datensammlung, also etwa in einer Datenbank oder im speziellen Falle in einem OPAC. So, wie man als Indexierer nach inhaltlicher Analyse eines Textes / Buches / Aufsatzes anhand eines Thesaurus zu einem "Deskriptor" (oder mehreren) gelangt, der stellvertretend für den Inhalt des Dokuments steht, gelangt ein Informationssuchender anhand des Deskriptors (und ggf. einer geeigneten Verknüpfung mehrerer Deskriptoren) zu dem durch diese(n) repräsentierten Dokument. Dabei kann (bzw. muss) er ggf. auch von den im Thesaurus verzeichneten Beziehungen Gebrauch machen, um seine sog. "Suchanfrage" zu präzisieren (auch um von einem Nichtdeskriptor zu einem Deskriptor zu gelangen usf.).

Ein Vorteil eines Thesaurus - so wird zumindest argumentiert, etwa gegenüber einem formalen Notationssystem - ist seine große Nähe zur "Sprache" des Benutzers (d.h. des Informationssuchenden), ein weiterer ist seine "Anpassbarkeit" an Veränderungen im Wissensbereich (Fachbereich).

Will man diesen Fachbereich ausweiten, etwa im Zusammenhang mit der Nutzung in einer Universalbibliothek, sind - u.a. wegen der semantischen Mehrdeutigkeit vieler Wörter, gerade auch der hochfrequenten Wörter einer Sprache - zusätzliche Maßnahmen nötig. Diese können (theoretisch) sein:

- die ausgiebige Differenzierung über Nichtdeskriptoren, Definitionen und Worterläuterungen;
- die Differenzierung von Deskriptoren durch natürlichsprachige oder nichtnatürlichsprachige Zusätze (etwa Bedeutungsnummern).

Aus formaler Sicht ist es damit bei entsprechender Differenzierung möglich, einen solchen Thesaurus auch "universell", d.h. mit Bezug zu allen Wissensgebieten der Menschheit, zu gestalten und einzusetzen. Voraussetzung ist es allerdings, dass man sich auf die (in der jeweiligen Sprache) verwendbaren Wörter (einerseits) und die Differenzierungstiefe (andererseits) verständigt.

4. Vom "Relationenwörterbuch" zum "Thesaurus"

Der Zusammenhang zwischen dem im Rahmen von IDX verwendeten "Relationenwörterbuch" und einem "klassischen" Thesaurus kann wie folgt beschrieben werden:

- Das Relationenwörterbuch dient in erster Linie der Unterstützung der automatischen Voll- oder Freitextindexierung. Hierbei werden die im Text (beim OPAC: Titel) vorkommenden Wörter auf Grundformen zurückgeführt, im Text (bzw. Titel) vorkommende Mehrwortbegriffe werden identifiziert; soweit es sich um Komposita handelt, werden auch Beziehungen zwischen den Teilwörtern und dem im Text / Titel stehenden Wort hergestellt.

- In zweiter Linie lassen sich im Relationenwörterbuch (teilweise intellektuell erschlossene) Beziehungen zwischen Derivationen herstellen (Beispiele: besuchen / Besuch, anfahren / Anfahrt, Begehung / begehen / begehbar ...).
- Zusätzlich lassen sich im Relationenwörterbuch Synonymbeziehungen abbilden und (beim Retrieval) nutzen. Beispielsweise lässt sich "Sonnabend" zu "Samstag" als Synonym zuordnen. Hierzu rechnen auch spezielle Synonymbeziehungen, etwa zwischen einer Abkürzung / einem Akronym und der entsprechenden Langform.
- Mit Hilfe der "Übersetzungsrelation" (oder auch IDX-spezifisch: anhand eines Übersetzungswörterbuchs) lassen sich schließlich auch Wörtern einer Sprache Übersetzungsäquivalente zuordnen.

Da das Indexierungsverfahren keine semantische Disambiguierung leistet, sind (zumindest bezogen auf die vollautomatische Erschließung) keine weiteren Relationierungen realisiert, obgleich das Relationenwörterbuch offen ist für die Markierung nahezu beliebiger Wortbeziehungen.

U.a. im Zusammenhang mit den Zielen von MILOS II eröffnet der nächste Entwicklungsschritt die Möglichkeit, Bedeutungsdifferenzierungen vorzunehmen. Das bestehende Beschreibungsinventar wurde dazu um folgende Möglichkeiten erweitert:

- Vergabe eines sog. "Semantik-Merkmals"
- Vergabe einer sog. "Bedeutungsnummer".

Hinzu kommen weitere Verfeinerungen im Wortklassenbereich, etwa zur Angabe von Vornamen, Körperschaftsnamen, Orts- und Städtenamen.

Das bestehende Relationenwörterbuch "kennt" bereits eine Reihe von Relationen, die für die Zwecke von MILOS II besonders interessant sind. Dazu gehören u.a. die folgenden Beziehungsangaben:

- Nichtdeskriptor => Deskriptor (und umgekehrt; oder analog: Stichwort -> Schlagwort)
- Unterbegriff => Oberbegriff (und umgekehrt),
- Wort => Klassifikation (und umgekehrt).

Mit Hilfe dieser Relationen und weiterer (bestehender) Verknüpfungen (wobei das Relationensystem im Grunde offen ist für weitere Verknüpfungen) soll auf der Basis der Schlagwortnormdatei der Deutschen Bibliothek (SWD) ein entsprechendes Netzwerk aufgebaut werden.

Die SWD differenziert dabei bereits anhand einer Identnummer die Vorzugsbenennungen für die Schlagwortvergabe. Dies gilt allerdings nicht für die "Oberbegriffs"-Verweise.

5. Aufbau eines SWD-Thesaurus, Anwendungsmöglichkeiten (Stand der Entwicklungen)

Nach der Entwicklung des Basiskonzepts für einen SWD-Thesaurus wird in einem nächsten Schritt eine Umsetzung der SWD-Daten in ein spezifisches SWD-Relationenwörterbuch erfolgen, um die notwendigen - vorwiegend der semantischen Differenzierung dienenden - Markierungen vorzunehmen. Dabei haben folgende Merkmale "differenzierende" Wirkung:

- Wortklassenangaben,
- semantische Merkmale,
- Bedeutungsnummern.

Eine Bedeutungsnummer sollte nur vergeben werden, wenn die Wortklassen- und Semantikmerkmale nicht ausreichend differenzieren.

Die Differenzierung wird allerdings nach Möglichkeit unter "universalen" Kriterien erfolgen. Dies bedeutet, dass ein entsprechendes Merkmal auch dann zu vergeben ist, wenn es innerhalb der "SWD-Welt" nur eine festgelegte "Bedeutung" gibt, einem Wort (- Eintrag) "außerhalb" der SWD-Anwendung jedoch mehrere Bedeutungen zugeordnet werden können. Diese Arbeiten betreffen nicht allein die SWD-Einträge, sondern im Prinzip alle Wörter der bearbeiteten Titel einschließlich der Teilwortrelationen. Da der Aufwand hier noch nicht zu überschauen ist, wird eine Anwendung zunächst nur experimentell erfolgen.

Neben diesen "lexikalischen" Arbeiten sind die technischen Verfahren bzw. Schnittstellen anzupassen. Dazu gehört die Entwicklung eines kontextbezogenen Selektionsverfahrens. Natürlich kann aufgrund des sehr eingegrenzten Kontext-Rahmens einer Titelangabe keine wirksame vollautomatische Lösung erwartet werden. Wie im Projektkonzept schon erläutert wurde, sollen daher experimentell auch schon existierende Notationen in das Erschließungs- und Disambiguierungsverfahren einbezogen werden.

6. Literaturübersicht:

Krause, Jürgen: PADOK: Test und Vergleich von Texterschließungssystemen für das deutsche Patent- und Fachinformationssystem. Regensburg 1986.

Kroupa, Edith: Strategien der Dokumentrepräsentation bei CTX. Ein Verfahren zur computergestützten Texterschließung und Textwiedergewinnung. In: I. Batori, S. Krause, H.-D. Lutz (Hrsg.): Linguistische Datenverarbeitung. Sprache und Information Bd. 4, Tübingen 1982, S. 155 - 166.

Lepsky, Klaus: Maschinelle Indexierung von Titelaufnahmen zur Verbesserung der sachlichen Erschließung in Online-Publikumskatalogen. Köln 1994.

Lustig, Gerhard (Hrsg.): Automatische Indexierung zwischen Forschung und Anwendung. Hildesheim 1986.

Salton, Gerald: Information Retrieval - Grundlegendes für Informationswissenschaftler. Hamburg, New York 1987.

Zimmermann, Harald H. (Hrsg.): Computergestützte Texterschließung mit CTX. Beiträge zum I. Forum Informationswissenschaft und Praxis. Veröffentlichung der Fachrichtung Informationswissenschaft. Saarbrücken 1983.

Zimmermann, Harald H.: Sprachsoftware für PC und Workstations. Saarbrücken 1993, 3 Bände (zu beziehen gegen Schutzgebühr über: SOFTEX GmbH, 66111 Saarbrücken, Schmollerstr. 31).

Anmerkungen

1 vgl. dazu und allgemein einführend: K. Lepsky: Maschinelle Indexierung von Titelaufnahmen zur Verbesserung der sachlichen Erschließung in Online-Publikumskatalogen, Köln 1994, bes. S. 23 ff.
2 Vgl. dazu ausführlich K. Lepsky, a.a.O., S. 32 ff.
3 Vgl. auch K. Lepsky, a.a.O., S. 38 f.