Rezensionen

Lenders, Winfried / Willée, Gerd:

LINGUISTISCHE DATENVERARBEITUNG – EIN LEHRBUCH.

Opladen: Westdeutscher Verlag, 1986. 201 Seiten.

Angesichts einer sehr differenzierten Angebots- und Bedarfssituation im Bereich der LDV an den bundesdeutschen Hochschulen (ausgebaute Studiengänge mit Tradition einerseits, Aufbauanfänge andererseits und schließlich noch Interesse-Bekundungen) sind in letzter Zeit mehrere Publikationen mit einführenden Intentionen erschienen, die jedoch alle jeweils andere Schwerpunkte setzen und sehr unterschiedlichen Ansprüchen und Erwartungen gerecht zu werden versuchen. Das Anwenderinteresse am Gebiet der LDV ist enorm gewachsen; es fehlte bislang jedoch an Einführungen, die diesen Titel verdienen. Dazu werden - eigene mühevolle Erfahrung ! - praxisorientierte Basisinformationen benötigt. Das Buch von *Lenders/Willée* wendet sich an Lehrende und Lernende, die zur Erforschung von Sprache und im Umgang mit Sprache die Hilfsmittel der Datenverarbeitung einsetzen wollen. Ihnen sollen die grundlegenden Verfahren der Verarbeitung von Texten thematisch und praktisch vermittelt werden,.."(S. 10) Als "Verarbeitung von Texten" ist aber offenbar nicht auch "Textverarbeitung" zu verstehen, da diese kurz vorher von der LDV abgegrenzt wird.

Im ersten Kapitel (Kommunikationstechnologie und Sprache) werden zunächst allgemeine theoretische Überlegungen zur Beziehung zwischen Sprache und moderner Kommunikationstechnologie angestellt. Von hier wird zur Maschinellen Übersetzung als einem Hilfsmittel der internationalen Kommunikation übergeleitet, deren drei Grundschritte (Analyse, Transfer, Synthese) in aller Kürze abgehandelt werden. In Kapitel 2 (Linguistische Grundlagen) werden Möglichkeiten des Computereinsatzes bei sprachlichen Analyseproblemen erörtert (z.B. Lemmatisierung, morphologische Operationen, Laut- und Silbensegmentierung). Mit der grundlegenden Methodik von maschineller Sprachverarbeitung (Algorithmen) beschäftigt sich Kapitel 3 (Verarbeitungstechniken); an Einzelbeispielen wird das Verfahren der Blockdiagrammdarstellung bei der Problemanalyse erläutert.

Die Vorteile des Einsatzes eines Computers in den in Kapitel 2 und 3 angesprochenen Problemfeldern werden in Kapitel 4 an relevanten Aufgaben deutlich gemacht, alles Arbeiten, die früher ohne Maschine nur mit großem, kostspieligem Arbeitsaufwand und der Gefahr der Fehlerhaftigkeit erledigt werden konnten.

Den entscheidenden Schritt zur Praxis linguistischer Arbeit mit dem Rechner will dann Kapitel 5 tun (Praktische Umsetzung von Algorithmen in Programme). Hier werden Grundstrukturierungen von Computerprogrammen und deren Verwendungsmöglichkeiten an interessanten Beispielen erläutert. Daß die Autoren sich dabei auf die Verwendung einer der möglichen Programmiersprachen festlegen (hier: PL/1), ist selbstverständlich, und die Auswahl ist auch begreiflich, da die Autoren Zugang zu einer IBM-Anlage haben. Diese Entscheidung ist dennoch bedauerlich, da PL/1 m.E. durchaus nicht so stark verbreitet und so leicht erlernbar ist, wie behauptet, vor allem aber weil diese Sprache auf anderen als IBM- Anlagen oft nicht optimal implementiert und daher mit größeren Schwierigkeiten bei ihrer Verwendung zu rechnen ist. Schade ist auch, daß sich die Autoren mit dem Lehrbuch zwar an Interessenten wenden, die "sich zum ersten Mal mit den Möglichkeiten der Datenverarbeitung" (S. 12) befassen, später jedoch eingestehen müssen, daß "Grundkenntnisse und eine gewisse Vertrautheit mit PL/1" vorausgesetzt werden. Einem Anfänger kann man das Buch daher als erste Informationsquelle nicht vorbehaltlos empfehlen; gleichwohl haben die Verfasser große Verdienste damit erworben, sich mit großer Sachkenntnis und interessanter Darstellung auf eine Expedition in begehrter werdendes Neuland begeben zu haben. Für eine Neuauflage, die dieses Buch sicherlich bald erhalten wird, wäre eine konsequentere Orientierung auf die angesprochene Zielgruppe wünschenswert.

Martina Schwanke, Universität Kiel

Gerhard Lustig (Hrsg.):

AUTOMATISCHE INDEXIERUNG ZWISCHEN FORSCHUNG UND ANWENDUNG

Linguistische Datenverarbeitung, Band 5. Hildesheim: Olms 1986. 182 Seiten.

In der GLDV-Reihe *Linguistische Datenverarbeitung* im Olms-Verlag ist der Band "Automatische Indexierung zwischen Forschung und Anwendung", herausgegeben von G. Lustig, erschienen. In diesem Band werden Vorgehen und Ergebnisse der Darmstädter Projekte WAI, AIR und AIR/PHYS, die seit 1977 zum Gebiet der automatischen Indexierung in Darmstadt durchgeführt wurden, zusammenfassend dargestellt.

Mit den Darmstädter Projekten ist es gelungen, einen Forschungsansatz von den ersten experimentellen Stadien bis zu einem System weiterzuentwickeln, das unter realen Arbeitsbedingungen eingesetzt wurde. Dieser Einsatz wurde in einem großangelegten Retrievaltest vorbereitet und konnte weiter forschend begleitet werden.

In dem Band sind in Beiträgen der jeweiligen Bearbeiter die einzelnen Komponenten beschrieben.

Im ersten Beitrag zur Konzeption des Darmstädter Verfahrens von G. Lustig werden die wesentlichen Faktoren genannt, die das Konzept in die Landschaft des Information Retrieval einordnen. Indexierung wird konsequent als Problematik der Behandlung von Massendaten, wie sie in den gewählten Anwendungsbereichen (realer Informationssysteme) FSTA und Physik anfallen, verstanden. Indexierung bedeutet damit den Übergang von dem Text englischsprachiger Referate zu einem vorgegebenen Deskriptorensystem. Diese Aufgabe wurde und wird auch im größeren Rahmen bis heute überwiegend intellektuell gelöst. Die erste wesentliche Komponente bei der Automatisierung der Indexierung ist die Erstellung eines Indexierungswörterbuches, das linguistische Aspekte (Anweisungen zur Identifikation von Grund- und Stammformen sowie Mehrwortgruppen), dokumentarische Aspekte (z. B. Zuordnung von Facetten zu Deskriptoren) und statistische Aspekte (Deskriptorenfrequenzen zur Signifikanzabschätzung) in sich vereinigt. Wesentlich ist die Gewinnung von sog. z-Relationen zwischen Textermen und Deskriptoren, die auf der Basis intellektuell indexierter Dokumente gewonnen und im Wörterbuch notiert werden. Die automatische Indexierung setzt eine Relevanzbeschreibung (z-Werte) der zu indexierenden Dokumente voraus, nach konkurrierenden Verfahren wird dann die Menge der zuzuordnenden Deskriptoren ausgewählt.

Gegenstand des zweiten Kapitels ist die Entwicklung von Indexierungswörterbüchern (Gewinnung von z-Werten, aufschlußreiche Zahlen zum partiellen Parsing und zur Formelbehandlung in großen Textmengen). Anschließend werden im dritten Kapitel die beiden (im folgenden Retrievaltest konkurrierenden) Indexierungsverfahren nach dem Booleschen und nach dem Polynomansatzverfahren dargestellt.

Die Beschreibung des Retrievaltests in Kapitel vier macht in der Bewertung der Ergebnisse dieses doch recht umfangreichen Retrievaltests die begrenzte Aussagefähigkeit der gängigen Bewertungsmaße deutlich. An einigen Punkten, wie z. B. der geringen Überlappung der intellektuell und der automatisch erschlossenen relevanten Antwortdokumentmengen, zeigt sich, daß der Bereich realer Retrievalpraxis und ihrer Bewertung noch keineswegs abschließend geklärt ist. Das Zusammenwirken von automatischer und intellektueller Indexierung, von Deskriptor- und Freitextsuche, die Berücksichtigung weiterer Faktoren wie Indexierungstiefe, verschiedener Benutzer- oder Fachgebietsaspekte weisen auf ein Arbeitsgebiet hin, das bisher – sicher aufgrund der Schwierigkeiten, die ein großangelegter Retrievaltest mit sich bringt – keineswegs voll erschlossen ist, jedoch ein breites Feld für weitere Untersuchungen darstellt.

N. Fuhr geht dem Aspekt des Ranking mit gewichteter Indexierung im anschließenden Abschnitt nach, die dargestellten Ergebnisse (Tanimoto und Cosinus-Maß vs. Fuzzy Retrieval und Extended Boolean Retrieval) wecken die Neugier auf seine vor kurzem abgeschlossene Dissertation, die auf den Daten des Retrievaltests beruht. Abschließend diskutiert G. Knorz die Möglichkeiten und Bedingungen des Einsatzes der automatischen Indexierung, die er (analog zur Entwicklung in anderen Bereichen der maschinellen Sprachverarbeitung) in einer intelligenten Kooperation mit dem (menschlichen) Indexierer sieht, der so von Routineaufgaben entlastet, Freiraum für (immer nötige) Konsistenzprüfungen und Adaptionen an neue Entwicklungen erhält.

Insgesamt bietet der Band eine zusammenhängende Dokumentation zu dem gesamten Darmstädter Indexierungsprojekt und zeigt Ergebnisse, Probleme und Dimensionen auf, mit denen unter realen Einsatzbedingungen zu rechnen ist.

Die diffusen Ergebnisse des Retrievaltests rühren möglicherweise von der Konstellation der verwendeten Indexierungsverfahren (Adaption des automatischen Verfahrens an der intellektuellen Indexierung) her. Bereits mit den im Band selbst angesprochenen Fragestellungen (s. o.) zeichnet sich jedoch die Möglichkeit ab, mittels weiterer analytischer Untersuchungen anhand der Daten des Retrievaltests Aufschlüsse über Indexierungs- und Retrievalprozesse zu gewinnen. Wie sich auch hier wieder gezeigt hat, ist eine Weiterentwicklung der Forschung auf diesem Gebiet auf der Grundlage globaler Bewertung kaum möglich; wohl nur durch die Analyse der unter realen / realitätsnahen Bedingungen ablaufenden Vorgänge ist ein Aufbrechen der Komplexität des Information Retrieval und damit eine Weiterentwicklung zu leisten.

Christine Schneider, Linguistische Informationswissenschaft: Regensburg