Rezensionen

J. Krause (Hg.):

Inhaltserschließung von Massendaten, Hildesheim, 1987

Krause, Jürgen (ed.): Inhaltserschließung von Massendaten. Linguistische Datenverarbeitung, Bd. 8. *Hildesheim, Zürich, New York: Olms,* 1987. 248 *S.*

Gegenstand der in diesem Band vorliegenden Untersuchung PADOK (Patentdokumentation) ist die inhaltsorientierte Erschließung von großen Textmengen, wie sie bei Freitextdatenbanken anfallen. Verschiedene linguistisch basierte Verfahren werden am Beispiel von Patentdatenbanken einem reinen Freitextsystem gegenübergestellt.

Der Band beschreibt die einzelnen Verfahren, die Testdurchführung und die Retrievalergebnisse auf der Grundlage von etwa 11.000 Patent-Abstracts und Patent-Titeln. Bewertungsparameter waren im Information-Retrieval verwendete Größen wie Recall und Precision mit geringfügigen Erweiterungen. Analysen auf dieser Grundlage wurden in PADOK in Relation zu Aufwandsanforderungen gesetzt.

Bei den vom Fachgebiet "Linguistische Informationswissenschaft" unter Prof. Krause (Universität Regensburg) in Zusammenarbeit mit dem Deutschen Patentamt, Mitarbeitern der GID, Systemanbietern und Industrie- und Fachinformationszentren untersuchten Verfahren handelte es sich um:

- Ein Freitextverfahren, das mit der üblichen Einzelwort-Segmentierung jedes Textwort in seiner im Text vorkommenden Form zum Deskriptor macht und invertiert.
- Das Verfahren PASSAT der Siemens AG, das für die Deutsche Sprache Komposita zerlegt und Vollformen auf Grundformen reduziert.
- Das Verfahren CTX der Universität Saarbrücken, das über die Funktionen von PASSAT hinaus die Möglichkeit bietet, zweigliedrige Mehrwortbegriffe aus größeren Einheiten zu extrahieren und für das Retrieval zur Verfügung zu stellen.

Zusätzlich wurde das Verfahren DETECT der GID ausschließlich im Rahmen der Erschließungsbewertung berücksichtigt. Dieses Verfahren isoliert Mehrwortbegriffe in ihrer im Text auftreten den maximalen Länge und bietet sie zur Indexierung und Recherche an.

Die Ergebnisse des Testes lassen sich wie folgt zusammenfassen: Läßt man Überlegungen zum Aufwand außer acht, sind die linguistisch motivierten Verfahren PASSAT und CTX dem Verfahren der Freitextindexierung überlegen. Da der Recall-Bewertung größeres Gewicht beigemessen wurde, und hier PASSAT bessere Ergebnisse als CTX lieferte, wurde PASSAT im Rahmen von PADOK als das erfolgreichere System bewertet. Bessere Precision-Werte bei CTX konnten diese Überlegenheit nicht ausgleichen.

Hinzu kommt noch ein um den Faktor 5 bis 12 schlechterer Wert bei der Aufwandsabschätzung bei CTX im Vergleich zu PASSAT. Die Aufwandsabschätzungen werden dankenswerterweise in der Untersuchung angegeben und es dürfte keine einfache Aufgabe sein, in einem weiteren Evaluierungsschritt in Zusammenarbeit mit den Anwendern die Aufwände in Relation zu den tatsächlichen Verbesserungen zu setzen. Sehr hohe Aufwände für Lexikon-Updates, die zu Lasten des Benutzers gehen und zum Teil sehr aufwendige manuelle Eingriffe während der Erschließung sind Gesichtspunkte, die insbesondere vor dem Hintergrund der Verwendung von Volltext-Datenbanken zu berücksichtigen sind.

In diesem Zusammenhang ist der Hinweis von Krause am Schluß des Berichts auf mögliche Vorteile von Verfahren zur Retrieval-, Indexierungs-, Thesaurusunterstützung, die zunächst einmal die rigiden Architekturen und die daraus resultierenden, eingeschränkten Indexierungs- und Recherchemöglichkeiten (Vollformen- Einzelwort-invertierung, Boole'sche Operatoren, Abstandsmaße) bestehender Freitextsysteme so belassen wie sie sind, recht sinnvoll. Eine derartige Philosophie entspricht im übrigen den Entwicklungen auf dem sehr weit entwickelten US Online-Markt, wo solche Verfahren als Zusatzfunktionen zu Systemen oder als Bausteine in solchen als "Expertensysteme für das Information Retrieval" oder pauschal als "Intelligentes Information Retrieval" in wissenschaftlichen Publikationen diskutiert und von Softwarehäusern entwickelt werden. Die Forschung und Entwicklung konzentriert sich hierbei angesichts eines Datenbankangebots, das weltweit zu 80 % aus englischen Datenbanken besteht, zur Zeit vorwiegend auf die Analyse der englischen Sprache.

Im Zuge des stetig anwachsenden bundesdeutschen Datenbankmarktes werden industrielle An-

wender (Anbieter, Bosts) auch für die deutsche Sprache Systeme mit einer Funktionalität anbieten, die der von PASSAT vergleichbar ist oder die über diese hinausgeht. Die derzeit vorwiegende Beschäftigung mit dem Englischen in der Information und Dokumentation hat weniger mit den zugegebenermaßen höheren Schwierigkeiten bei der linguistischen Analyse des Deutschen im Gegensatz zum Englischen zu tun als mit der Tatsache, daß ein Online-Markt in der BRD, insbesondere was die Produktion von deutschsprachigen Datenbanken anbetrifft, erst im Entstehen ist.

Tatsächlich bietet auf dem US-Markt z.B. BRS seit einiger Zeit eine linguistische Indexierungs- und Retrievalkomponente an, während sich STATUS über die Verwendung vorwiegend statistischer Funktionen immerhin "intelligenten" Retrievalfunktionen nähert und diverse englische und US-Firmen linguistische Software-Tools im Bereich der Information und Dokumentation anbieten. Es ist zu erwarten, daß gerade auf dem Gebiet der Massentextverarbeitung die Computer-Linguistik in den nächsten Jahren zu Fortschritten beitragen wird.

Mit dem Bericht zum PADOK-Test liegt eine für den in der Information und Dokumentation Beschäftigten außerordentliche Untersuchung vor: Eine Untersuchung gegen den Strom geschrieben, weil in großem Stil Praktiker aus Industrie und Anwendung miteinbezogen wurden und weil anhand realer Daten getestet wurde. Eine Untersuchung, die gegen den Mainstream computerlinguistischer Forschung aufzeigt, daß letztere noch weit von realen Anwendungen in der Verarbeitung von Massendaten entfernt sind. Allerdings auch eine Untersuchung, die zeigt, daß in Bereichen großer innovativer Erstarrung wie den Bereichen der Indexierung (und damit auch der Recherche) bald mit neuen Ansätzen zu rechnen ist.

Das Buch zeigt auch auf, in welch hohem Maße noch Evaluierungsarbeit gerade im "klassischen" Retrieval zu leisten wäre. Es gibt z.B. so gut wie keine umfassende Studie zu den Vor- und Nachteilen beim Einsatz von Kontextoperatoren. Wenn eine solche vorliegt, wird die Frage des Einsatzes von weitergehenden linguistischen Verfahren insbesondere im Bereich der Syntax besser zu führen sein als derzeit. Die Untersuchung von Bauer aus dem PADOK-Team bringt für den Bereich der Trunkierung hochinteressante Ergebnisse, die auf die Nützlichkeit linguistisch morphologischer Analysen hinweisen.

Kritisch wäre anzumerken: Derjenige, der an der Anwendung in der Praxis und damit an den wirtschaftlichen Aspekten derartiger Verfahren interessiert ist, hätte sich eine prägnantere Zusammenfassung der Ergebnisse gewünscht ebenso wie zusätzliche Schaubilder zu den Ergebnissen. Auch eine bessere Vergleichbarkeit der Zeit aufwände über eine Relativierung der unterschiedlichen Rechnerleistungen für die Erschließungsläufe wäre wünschenswert gewesen.

Insgesamt: Eine lohnende Lektüre für den Fachmann in Industrie und Universität! Informationen zu

BRS: BRS Europe, 11 Weymouth Street, London WIN 3FG

PASSAT: SIEMENS, DV 7, Otto-Bahn-Ring 6, 8000 München 83

STATUS: STATUS Marketing, Barwell Computer Power Ltd. Curie Avenue, Barwell Oxfordshire OXll OOW

Christoph Schwarz, Siemens AG, DI AP 323

Alexander Roßnagel, Peter Wedde, Volker Hammer, Ulrich Pordesch:

"Die Verletzlichkeit der Informationsgesellschaft",

Westdeutscher Verlag Opladen, 1989.

Kurzkommentar:
Eine facettenreiche Studie der Entwicklungen, Rahmenbedingungen und Folgen der Informations- und Kommunikationstechniken unter besonderer Berücksichtigung der Vernetzung, deren Eigengesetzlichkeiten (Sicherungserfordernisse) in der Sicht der Autoren Einschränkungen von Freiheit und Demokratie befürchten lassen. Mit ihren pointierten Aussagen bietet die Schrift Stoff zu fruchtbaren - Kontroversen.

Newsletter Technologiefolgenabschätzung, Jahrgang 1-1989, *Nr. 0*