NATÜRLICHSPRACHIGE PROBLEMBESCHREIBUNG ALS EIN VERFAHREN FÜR DEN BÜRGERNAHEN ZUGANG ZU DOKUMENTATIONSSYSTEMEN

Harald H. Zimmermann

Regensburg

Abstrakt

Die rasche Entwicklung der neuen Kommunikations-Medien die Bildschirmtext, Zweiweg-Kabelfernsehen, aber auch der preiswerte Zugang zu Informationssystemen über Paketvermittlungsnetze eröffnet zunehmend die Möglichkeit der Anbindung weiter Bevölkerungskreise an Informations- und Dokumentationssysteme.

Die Computerlinguistik, besonders die Grundlagenforschung im Bereich der Künstlichen Intelligenz, beschäftigt sich seit vielen Jahren bei zunehmender Tendenz mit Fragen der Repräsentierung und Erschließung von in natürlichsprachiger Form gespeichertem Wissen. Dies geschieht einerseits in der Absicht, Erkenntnisse über die Funktion und das Funktionieren von Sprache zu gewinnen, zu vertiefen und zu erproben, andererseits aber auch in der Absicht, solche Verfahren - etwa im Rahmen von Frage-Antwort-Systemen - (zumeist modellhaft) in Anwendung zu bringen.

Derartige Verfahren sind jedoch derzeit nur für sehr enge Themenbereiche - und auch hier nur mit Einschränkungen - anwendbar. Eine Ausdehnung auf größere 'Welten' oder Weltausschnitte scheitert zumindest an dem großen intellektuellen Aufwand, der für eine entsprechend tiefe und umfassende Wissensaufbereitung erforderlich ist.

Dem gegenüber steht inzwischen ein erheblicher Bedarf an Informationssystemen, die vom ungeübten Bürger ohne größere technische und formale Schwierigkeiten bedient werden können, also nicht - wie bisher - einen Systemspezialisten als Vermittler einschalten. Dazu bietet sich heute u.a. eine Bedienerführung über den sog. graphischen Dialog oder über die Menütechnik als Auswahl- und Entscheidungsverfahren an. Derartige Verfahren sind jedoch nicht immer ausreichend flexibel und zudem langwierig.

Eine Alternative im Bereich des Referenzretrieval stellt die natürlichsprachige Problembeschreibung dar, wie sie im Regensburger System JUDO (für 'juristische Dokumentbeschreibung', hergeleitet aus dem Anwendungsbereich (Datenschutz-Recht)) integriert ist: Die 'Suchfrage" beim Dokument-Retrieval besteht aus einem oder mehreren (Teil-) Sätzen.

Die Sätze werden über linguistische und z. T. probabilistische maschinelle Verfahren in der gleichen Weise bearbeitet, wie zuvor die Dokumente der Datenbank. Auf diese Weise wird zugleich eine Homogenisierung der Dokumenterschließung (Indexierung) und der Dokumentidentifikation (Retrieval) erreicht: Dazu werden u.a. Paraphrasierungen/ Normierungen vorgenommen und Thesaurusrelationen herangezogen.

Während der Textanalyse wird versucht, syntaktische und semantische Mehrdeutigkeiten aufzulösen. Dazu wird u.a. die sog. 'Saarbrücker Automatische Textanalyse (SATAN) verwendet; die Informations-Retrieval-Systeme TELDOK und GOLEM dienen z. Zt. als Implementierungsgrundlage der Retrieval-Komponente. Folgende Fragen werden im Referat - belegt durch Beispiele und Statistiken - behandelt:

- o Texterstellung und -aufbereitung, Textbasis, Dokumenttypen;
- o Allgemeinsprachliche und fachsprachliche maschinelle Analyse;
- o Deskriptoren (einfache/komplexe Deskriptoren);
- o Thesauruserstellung und -relationen;
- o Natürlichsprachliche Problembeschreibung, besonders Probleme der Paraphrasierung.

Den Abschluß bildet ein Ausblick auf noch zu lösende Fragen und auf Anwendungsmöglichkeiten, auch im Hinblick auf die Übertragbarkeit auf andere Themenbereiche.