Das KONTEXT–Textanalysesystem

Karin Haenelt
IPSI, Darmstadt

Am Institut für Integrierte Publikations- und Informationssysteme (IPSI) der GMD, Darmstadt wurde unter Leitung von Frau Dr. Karin Haenelt das KONTEXT–Textanalysesystem konzipiert und entwickelt, das u.a. auch auf der Buchmesse 93 (Frankfurt) vorgestellt wurde.

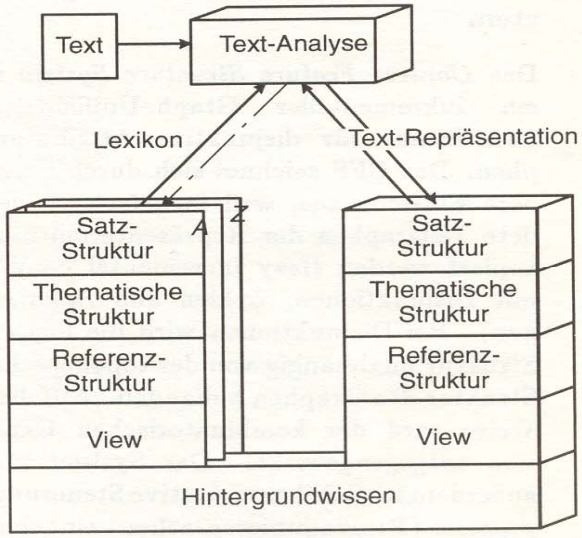
Das KONTEXT–System ist ein erster Prototyp einer neuen Generation von Textanalysesystemen. Grundlage des Systems ist das im IPSI neu entwickelte KONTEXT–Textmodell, das es ermöglicht, Texte in völlig neuartiger Weise textorientiert (nicht auf der Basis isolierter Wörter oder einzelner Sätze) unter Berücksichtigung von Textstruktur und Kontext zu verarbeiten. Dank dieses Modells können in der Textanalyse auch in Texten vorkommende neue Informationen erkannt und integriert werden.

Wissenschaftlicher Hintergrund

a) Das KONTEXT–Modell

Das KONTEXT–Textmodell beschreibt, wie mit Mitteln der natürlichen Sprache (neues) Wissen aufgebaut und mitgeteilt wird. Die Grundannahmen des Modells sind

- die Information, die in einem Text vermittelt wird, sowie die Information, die die kontextuelle Organisation dieser Information beschreibt, kann in fünf Ebenen strukturiert werden: Satzstruktur, Thematische Struktur, Referenzstruktur, View (textspezifische Sicht auf Wissen), Hintergrundwissen.
- der Grundmechanismus der Konzeptkonstruktion ist der Diskurs. Diskurs ist definiert als eine Sequenz von Zustandsübergängen zwischen Diskursstates, und Diskursstates sind definiert durch die Information, die in den fünf Ebenen repräsentiert ist.



Der Textanalyseprozeß konstruiert und durchläuft die Ebenen unter der Kon-

trolle der Diskursentwicklung. Auf diese Weise kann er inkrementell die text spezifische Sicht (View) auf Hintergrundwissen aufbauen.

b) Die Textgrammatik

Eine lexikalisierte Textgrammatik beschreibt den Beitrag, den sprachliche Ausdrücke zum Aufbau der Schichten des KONTEXT-Modells leisten. In der Entwicklung dieser Textgrammatik wird u. a. auch der Zusammenhang zwischen Tiefenkasus (semantischen Rollen) und syntaktischen Funktionen (Satzteile, Kasus) systematisiert. Alle Schichten (also nicht nur die Syntax, sondern auch die Wissensrepräsentation) werden in einem einheitlichen Formalismus, dem *Context Feature Structure System* (CFS) repräsentiert. Das Analyseverfahren erzeugt schrittweise lesend eine linguistische und konzeptuell angereicherte Textrepräsentation. Im Sinne des Integrierten Parsing werden alle Beschreibungsebenen des Modells zum Zwecke der wechselseitigen Ergänzung und Einschränkung gleichzeitig konstruiert.

c) Das Context Feature Structure System

Das *Context Feature Structure System* ist ein inkrementeller Graph-Unifikationsformalismus für disjunktive Attributgraphen. Das CFS zeichnet sich durch besondere Effizienz aus, weil mehrfach verwendete Teilgraphen der Repräsentation nicht kopiert werden (lazy incremental copying mit Disjunktionen, Zyklen und Rekursionen). Bei Disjunktionen wird die logische Struktur unabhängig von der topologischen Struktur des Graphen behandelt. Auf diese Weise wird der kombinatorischen Explosion entgegengewirkt. Das System kann außerdem in nichtkommutative Steuerungsprozesse (Programmiersprachen) eingebunden werden, ohne daß die jeweiligen Datenstrukturen vollständig kopiert werden müssen.

Eine Lexikon- und Grammatikentwicklungsumgebung steht zur Verfügung.

Anwendungsgebiete

Der Vorteil der textmodellbasierten Verarbeitung natürlichsprachiger Texte (wie sie in KONTEXT realisiert ist) besteht darin, daß *ein und dasselbe Prinzip* zu einer *Vielzahl unterschiedlicher Anwendungen* führt. Kern ist die Textanalysekomponente. Sie bleibt für verschiedene Anwendungskontexte in ihrer Funktionalität jeweils die gleiche.

Die Textanalysekomponente baut eine mehrschichtige Textrepräsentation auf. Eine dieser Schichten enthält die aus Texten gewonnenen Fakten, die anderen enthalten die Textstruktur. Die Textrepräsentation erfüllt mehrere Funktionen:

=> sie ermöglicht einen direkten inhaltsorientierten Zugriff auf Volltexte, wobei der Zugriff durch Textinhalt und Textstruktur bestimmt wird (Textretrieval++);

=> sie kann mit Methoden des Faktenretrieval ausgewertet werden (Information Retrieval++);

=> sie kann als Thesaurus ausgewertet werden;

=> sie dient der automatischen Indizierung und Verschlagwortung von Texten;

=> sie dient als Grundlage Syntax- und Konzept-basierter Rechtschreibprüfungen;

=> sie bildet eine abstrakte Schnittstelle zu Texten und liefert so die Grundlage für weitere textuelle Operationen, wie automatische Paraphrasengenerierung (Lexikonaufbau), Maschinelle Übersetzung, Kondensierungen (Inhaltsangaben, Zusammenfassungen), Dialog-Anwendungen oder Aufbau, Analyse und Linearisierung von Hypertexten.

Ein weiterer Vorteil der textmodellbasierten Verarbeitung natürlichsprachiger Texte in KONTEXT besteht darin, daß auch *neues Wissen* systematisch behandelt werden kann. So lassen sich solche Systemkomponenten auch als „lernende Wörterbücher" und „lernende Faktendatenbanken" einsetzen, die neue Veröffentlichungen lesen, das in Texten mitgeteilte neue Wissen erkennen, es in ihre Textrepräsentation übernehmen und sich so selbst auf dem laufenden halten.

Kontaktadresse

Dr. Karin Haenelt
Institut für Integrierte Publikations- und Informationssysteme (IPSI)
Dolivostraße 15
6100 Darmstadt, Tel.: 869–828, e–mail: haenelt@darmstadt.gmd.de