Informatik-Fachberichte 243

Herausgeber: W. Brauer
im Auftrag der Gesellschaft für Informatik (GI)

Subreihe Künstliche Intelligenz

Mitherausgeber: C. Freksa
in Zusammenarbeit mit dem Fachbereich 1
„Künstliche Intelligenz" der GI

Udo Hahn

Lexikalisch verteiltes Text-Parsing

Eine objektorientierte Spezifikation eines Wortexpertensystems auf der Grundlage des Aktorenmodells

Springer-Verlag
Berlin Heidelberg New York London
Paris Tokyo Hong Kong Barcelona

Autor

Udo Hahn
Universität Passau
Fakultät für Mathematik und Informatik
Postfach 2540, W-8390 Passau

CR Subject Classification (1987): I.2.7, I.2.4, H.3.1

CIP-Titelaufnahme der Deutschen Bibliothek.
Hahn, Udo:
Lexikalisch verteiltes Text-Parsing: eine objektorientierte Spezifikation eines Wortexpertensystems auf der Grundlage des Aktorenmodells / Udo Hahn. – Berlin; Heidelberg; New York; London; Paris; Tokyo; Hong Kong; Barcelona: Springer, 1990
(Informatik-Fachberichte; 243: Subreihe künstliche Intelligenz)
Zugl.: Konstanz, Univ., Diss., 1987

ISBN-13: 978-3-540-53230-9 e-ISBN-13: 978-3-642-76132-4

DOI: 10.1007/978-3-642-76132-4

NE: GT

2145/3140-543210 – Gedruckt auf säurefreiem Papier

Vorwort

Bei der vorliegenden Arbeit handelt es sich um eine überarbeitete und aktualisierte Version meiner Dissertation, die im Februar 1987 von der Sozialwissenschaftlichen Fakultät der Universität Konstanz angenommen wurde. Zusammen mit der ebenfalls in den Informatik-Fachberichten (Bd. 198) veröffentlichten Dissertation von Ulrich Reimer dokumentiert die Arbeit den methodischen Entwicklungsstand der an Textverstehensproblemen orientierten informatischen Komponente des Faches Informationswissenschaft, in dessen institutionellem Rahmen beide Monographien entstanden sind. Der engere Projektkontext wurde durch das vom BMFT von 1981 bis 1986 geförderte Textkondensierungssystem TOPIC definiert, ein System zur flexiblen Zusammenfassung (Abstracting) deutschsprachiger Fachtexte aus dem Bereich der Informationstechnik.

Ausgangspunkt dieser Arbeit war das Problem, ein authentischen Fachtexten angemessenes linguistisches Beschreibungssystem für sprachliche Phänomene und ein Verfahren für ihre inhaltliche Analyse zu entwickeln. Zur Lösung dieses Problems ist eine Textgrammatik und ein entsprechender Text-Parser erarbeitet worden, die auf einem semantischen Beschreibungsansatz beruhen.

Eine Betrachtung der grundlegenden Merkmale semantischer Grammatiken zeigt, daß sie im wesentlichen Lexikon-Grammatiken sind; d.h. das gesamte für das Parsing relevante Wissen ist den lexikalischen Einheiten der Sprache zugeordnet. Zugleich sind semantische Grammatiken/Parser durch eine stark auf Kommunikation ausgerichtete Architektur gekennzeichnet - damit ist einerseits der parallele Zugriff auf unterschiedliche Wissensquellen beim Textverstehen gemeint, andererseits die aus der lexikalischen Verteilung erwachsende Notwendigkeit zu intensiven Interaktionen zwischen lokalen Parse-Prozessen bei der Interpretation zusammenhängender Strukturen auf der Phrasen-, Satz- und Textebene.

Im Bereich semantischer Grammatiken/Parser stellt das Wortexperten-Modell den konzeptionell innovativsten Versuch dar, die oben angesprochenen Merkmale in ein Sprachverstehensmodell zu integrieren. Gravierende methodische Mängel und nicht haltbare linguistische Hypothesen verlangten jedoch eine Reformulierung der ursprünglichen Wortexperten-Konzeption. Die wesentlichen in dieser Arbeit entwickelten Vorschläge können wie folgt zusammengefaßt werden:

Das hier vorgestellte Konzept eines lexikalisch verteilten Text-Parsers ist vollständig in einer objektorientierten Spezifikationssprache auf der Grundlage des Aktorenmodells beschrieben. Darauf aufbauend

a) werden die deklarative Struktur der Textgrammatik und das prozedurale Verhalten des Text-Parsers einheitlich und auf einem gleichen Abstraktionsniveau behandelt,
b) sind die Sprachprimitive der Aktoren-Spezifikationssprache im Sinne des Konzepts abstrakter Datentypen formalisiert,
c) ist insbesondere die linguistische Spezifikation der semantischen Grammatik von der Spezifikation der Repräsentationsstrukturen des Welt- und Domänenwissens logisch strikt getrennt.

Im Mittelpunkt der linguistischen Spezifikationen stehen Beschreibungen generalisierbaren Verhaltens lexikalischer Objekte:

a) Für das Verhalten lexikalischer Klassen repräsentative Wortexperten-Prototypen werden nach funktionalen Gesichtspunkten formuliert und lösen das alte Wortexperten-Paradigma von der Beschreibung des idiosynkratischen Verhaltens individueller Wörter ab (funktionale Wortexperten-Prototypen kann man sich etwa als hochgradig reguläre Anaphora-, Ellipsen- oder Koordinationsgrammatiken vorstellen).
b) Prototypen setzen sich ihrerseits aus Subexperten zusammen. Ihre Spezifikation trägt zur Faktorisierung gemeinsamen Wissens in lexikalisch verteilten Grammatiken bei und reduziert erheblich die Beschreibungsredundanz in und das Wachstum von Wortexperten-Kollektionen.

Zur Bewältigung des grammatikalischen Komplexitätsproblems, dem sich alle Entwickler von natürlichsprachlichen Grammatiken mit dem Anspruch möglichst breiter linguistischer und konzeptueller Abdeckung gegenüber sehen, sind an die Körnung der Grammatikspezifikation bewußt solche Anforderungen gestellt worden, die auf einen partiellen semantischen Text-Parser für das Deutsche zielten. Das Beschreibungsproblem konnte somit darauf reduziert werden, die textuellen Bezüge von durch Nomen denotierten Konzepten auf einem "flachen" Verstehensniveau zu erfassen, wie es für die Erfüllung der Funktion des Textverstehenssystems TOPIC (Bereitstellung indikativer, d.h. themenbezogener Textzusammenfassungen), in das der Parser integriert ist, hinreichend ist.

Auf der Grundlage dieser methodisch motivierten Fortschritte läßt sich die hier gegebene Beschreibung als ein Beitrag zu einer formalisierten Theorie des semantischen Text-Parsing charakterisieren. Art und Konzeption der Beschreibungssprache garantieren zudem einen hohen Grad an Transportabilität in andere Domänen, was für semantische Grammatiken (wegen fehlender Abstraktionsmechanismen) bislang eher ungewöhnlich ist. Insgesamt stehen im Rahmen des hier entwickelten Beschreibungsansatzes konzeptuell hohe Mechanismen zur Verfügung, mit denen der Entwurf und die Wartung semantischer Grammatiken/Parser unter strenger planerischer Kontrolle bleibt.

Die Gültigkeit der textlinguistischen Beschreibungen beruht auf den Erfahrungen, die mit der implementierten Version dieser Textgrammatik im Rahmen einer insgesamt fünfjährigen Phase von Textkondensierungsexperimenten mit dem Textverstehenssystem TOPIC gesammelt wurden. Neben der Formalisierung der Textbezüge ist die hier beschriebene Textgrammatik also - wenn auch nicht unter Zugrundelegung strenger experimenteller Maßstäbe - in ihren Grundzügen validiert.

Wie immer bei längerfristigen Entwicklungsarbeiten in größeren Teams gibt es eine Reihe von Personen und sachlichen Umständen, die auf individuelle Arbeiten bestimmend einwirken. In meinem Fall gilt dies zu allererst für Prof. R. Kuhlen, der den institutionellen Rahmen für informationswissenschaftlich definierte Forschung aufgebaut und die wesentlichen Forschungsfragen thematisiert hat, die durch das TOPIC-Projekt konstruktiv ausgearbeitet wurden. Mit ausschlaggebend für die dabei erzielten Fortschritte war die Gewährung und Absicherung wissenschaftlicher Freiräume, in denen die TOPIC-Gruppe im Rahmen der Projektdefinition auf eine auch heute noch eher unübliche selbstbestimmte Weise forschen konnte. Nicht zuletzt diesen Rahmenbedingungen verdanken die beteiligten Wissenschaftler das durch eigene Publikationen erreichte Ausmaß an individueller Profilierung. Eine naturgemäß andere Form der Kooperation entwickelte sich bei der Zusammenarbeit mit meinem unmittelbaren Projektpartner Ulrich Reimer. Wir haben gemeinsam

die konzeptionellen und technischen Rahmenbedingungen des TOPIC-Systems entwickelt und waren folglich auch die "ersten" Diskussionspartner für fachliche Probleme im Zusammenhang mit unseren Dissertationen. Ulrich Thiel, Mitarbeiter in der TOPOGRAPHIC-Gruppe am Lehrstuhl, hat mich besonders in der Schlußphase der Fertigstellung dieser Arbeit durch viele ins Detail gehende Gespräche nachhaltig unterstützt. Prof. E. Pause, dem Korreferenten der Dissertation, verdanke ich wertvolle Hinweise auf die linguistische Signifikanz der hier berichteten Ergebnisse. Prof. M. Jarke ermöglichte mir schließlich die Fertigstellung der jetzt vorliegenden Version meiner ursprünglichen Dissertation in der neuen Passauer Arbeitsumgebung. Bei allen möchte ich mich aufrichtig für ihre Unterstützung bedanken. Dies gilt auch für die wissenschaftlichen Gesprächspartner auf Kongressen und Arbeitstreffen, die mein Verständnis des Wortexperten-Konzepts durch ihre Kritik und Anregungen herausgefordert und damit auch fortlaufend verschärft haben.

Passau, im Februar 1990 Udo Hahn

Inhaltsverzeichnis