Information Retrieval und Computerlinguistik

SIFT - Selecting Information From Text

Ein Projekt an den Universitäten von Heidelberg, Limerick und Amsterdam sowie von Lotus Development.

Dauer: 24 Monate

Aufwand: 123 Personen-Monate

Projektbeteiligte

- University of Limerick (Koordinierung)
- Lotus Development, Ireland
- University of Amsterdam
- Universität Heidelberg

Kontakt
Dr Richard Sutcliffe
Dept. of Computer Science
and Information Systems
University of Limerick
Limerick, Ireland
Telephone:
+353 61 333644 Ext. 5006
+353 61 330876 Fax

Ziele

Das SIFT Projekt erarbeitet den Prototypen eines intelligenten Online-Hilfesystems für Software Manuals. Zwei Konzepte bilden die Grundlage:

- Das Vektormodell für Information Retrieval
- Verteilte Muster, mit denen Textbedeutung eingefangen werden soll.

In seinem Endausbau soll der Prototyp eine Anfrage in natürlicher Sprache entgegennehmen, mit der ein Nutzer nach softwarerelevanten Aspekten sucht. Ergebnis wird eine nach vermuteter Relevanz geordnete Liste von Verweisen auf Textstellen sein (Ranking), die zur Klärung der Frage beitragen könnten. Das Projekt soll darüber hinaus zeigen, daß verteilte Muster effektiv in praktischen Sprachverarbeitenden Systemen eingesetzt werden können und daß eine Integration mit bestehenden Techniken und Systemen fur lexikalische Datenbanken und für robustes lexikalisches Parsing erfolgreich ist. Dies wurde im Prinzip bereits in einem praktisch eingesetzten Retrievalsystem mit Erfolg realisiert, wo man das Vektormodell mit verteilten semantischen Repräsentationen kombinierte. Allerdings wurde dieses System weitgehend manuell aufgebaut, und SIFT soll die Technologie dafür bereitstellen, solche Systeme automatisch zu erzeugen.

Ansatz

Das SIFT -System wird aus zwei Hauptkomponenten bestehen:

- Die Dokumentenanalyse wird SGML-strukturierte Manuals auf der Ebene von Kapiteln, Unterkapiteln und einzelner Sätze mit verteilten Mustern in Verbindung bringen, mit denen die Bedeutung der Texteinheiten repräsentiert werden soll.
- Die interaktive Query-Komponente wird die Anfrage entgegennehmen und aus dem Anfrageergebnis ein Ranking für Textstellen erzeugen.

Das System basiert auf der Anwendung robuster lexikalischer Parsing-Techniken, der Zuordnung semantischer Rollen zu syntaktischen Konstituenten und der Extraktion verteilter Repräsentationen aus maschinenlesbaren Wörterbüchern.

Praktischer Nutzen und Ausblick

Auf textuelle Information gezielt zuzugreifen ist ein Problem in jeder Organisation. Genau für diese Aufgabe demonstriert SIFT eine innovative Lösung. Darüber hinaus können und sollen die eingesetzten Techniken in ei

nem weiten Spektrum verwandter Aufgaben der Textbearbeitung eingesetzt werden. So wird an die Integration in Stil-Checker, Zusammenfassungs-Generatoren und in Werkzeuge zur computerunterstützten Übersetzung gedacht. Aber es werden auch theoretische Einsichten in die Anwendbarkeit von Techniken zur automatischen Textanalyse erwartet, z.B. inwieweit verteilte Repräsentationen tatsächlich zur gleichzeitigen Abdeckung von Wort- und Satzbedeutung benutzt werden können, inwieweit ein großes, robustes und nicht spezialisiertes semantisches Lexikon automatisch aufgebaut werden kann, und ob ein robustes lexikalisches partielles Parsen möglich ist. Teil des Projektes ist es, die Überführung der SIFT - Technologie in eine kommerzielles Produkt zu untersuchen.
Quelle: http://itdsrvl.ul.ie/SIFT/sift-home-page.html

Venia "Computerlinguistik" für Dr. Karin Haenelt

Habilitationsverfahren an der Universität Heidelberg erfolgreich abgeschlossen

Karin Haenelt (GMD, Darmstadt), hat bei der Neuphilologischen Fakultät der Universität Heidelberg eine Habilitationsschrift mit dem Titel "Das KONTEXT-Modell. Verarbeitung natürlichsprachiger Texte auf der Basis eines Textmodells" eingereicht und am 8. Mai 1996 "mit Bravour das Habilitationskolloquiuum bestanden" (p. Hellwig).

Eine schriftliche Ausarbeitung ihres Vortrags *"Nachschlageverfahren im (elektronischen) Bedeutungswärterbuch* " soll unter dem Titel *"Looking-Up Procedures in an Electronic Meaning Dictionary: Considerations on the Role of a Meaning Dictionary in Textual Communication"* in Lexicographica erscheinen.

Eine Kurzfassung der Arbeit ist zugänglich als Arbeitsbericht der GMD: *"Das KONTEXT-Modell und die Konzeption der textmodellbasierten Verarbeitung natuerlich-sprachiger Texte.*" (Nr. 1009. GMD: Sankt Augustin, Juli 1996).

Eine allgemeine - nicht in allen Aspekten ganz aktuelle Beschreibung - des Kontext-Projektes und -Modells findet sich unter

lnt: http://www.darmstadt.gmd.de/KONTEXT/kontext.html

In diese Beschreibung sind verständlicherweise auch Informationen aus der Habilitationsschrift eingeflossen (z.B. „Das Kontext-Modell").

Karin Haenelt verfugt nun über die Venia "Computerlinguistik", und das LDV-Forum gratuliert herzlich.