COMPASS - ein intelligentes Wörterbuchsystem für das Lesen fremdsprachiger Texte

Feldweg, Helmut

Historical Social Research, Vol. 22 — 1997 — No. 2, 256-262

HUMANITIES COMPUTING

COMPASS. Ein intelligentes Wörterbuchsystem für das Lesen fremdsprachiger Texte

*Helmut Feldweg (Tübingen)**

Mit der Einführung elektronischer Wörterbücher wurde das zeitaufwendige Nachschlagen erheblich vereinfacht. Das gilt vor allem dann, wenn auch der zu lesende Text in elektronischer Form vorliegt. Diese Lesekonstellation gewinnt mit der zunehmenden Ausbreitung von Computernetzen, elektronischen Büchern und Dokumenten immer mehr an Bedeutung.

Die elektronischen Wörterbücher selbst wie auch die Nachschlagetechnik werden jedoch den Möglichkeiten des elektronischen Mediums derzeit nicht gerecht. Bei den Wörterbüchern handelt es sich um eine elektronische Abbildung der als Druckmedium für den menschlichen Gebrauch konzipierten Nachschlagewerke. Die Nachschlagetechnik beschränkt sich im allgemeinen darauf, eine Zeichenkette im Text mit den Zeichenketten der Wörterbuchschlagworte zu vergleichen und bei einer Übereinstimmung den zugehörigen Eintrag auszugeben. Die intellektuellen Leistungen eines menschlichen Wörterbuchbenutzers werden von diesen Systemen nicht übernommen. Die Zurückführung von flektierten Formen auf deren Grundform, die Wortartbestimmung und die Auswahl der passsenden Bedeutung in einem längeren Wörterbucheintrag müssen weiterhin vom Benutzer geleistet werden.

Das Projekt COMPASS hat demonstriert, daß diese Einschränkungen herkömmlicher elektronischer Wörterbücher durch den Einsatz verfügbarer Techniken überwindbar sind. Dazu wurde ein Prototyp eines Computerprogramms entwickelt, der qualitativ hochwertige, strukturell aufbereitete Wörterbücher durch ein intelligentes, kontextsensitives Nachschlageverfahren erschließt und die Informationen dem Benutzer über eine ansprechende graphische Schnittstelle präsentiert

Die Leistung des Prototyps wurde durch eine erste Serie von Benutzertests beurteilt. Dabei wurde das System von den Testpersonen ausgesprochen positiv bewertet Einige Benutzerkommentare auf die Frage, ob COMPASS effizienter ist als ein Papierwörterbuch:

*** Protokoll des 67. Kolloquiums über die Anwendung der EDV in den Geistes Wissenschaften an der Universität Tübingen am 6. Juli 1996.**

- *»I get fed-up with leafing through paper dictionaries. I prefer being able to scan ahead like this.«*
- *»Chief advantage is the speed, and that the user can go on working on the text whilst Compass is accessing the translation options.«*
- *»More efficient particularly for a weak language competence.«*

Die Ergebnisse zeigen, daß das Lesen fremdsprachiger Texte durch ein System wie COMPASS erheblich vereinfacht wird und ein besseres Verständnis der Texte erreicht werden kann. Tatsächlich glauben wir, daß in vielen Fällen, wenn der Leser bereits über Grundkenntnisse der Fremdsprache verfügt, das Übersetzen von Texten durch den Einsatz eines solchen Systems vermieden werden kann.

In den nachfolgenden Abschnitten werden die Komponenten des Prototyps und die Organisation des COMPASS-Projekts detaillierter beschrieben.

Die Wörterbücher

Die lexikographische Grundlage des Projekts bilden das *Collins-Klett-Großwörterbuch Deutsch-Englisch* und das *Oxford-Hachette-Wörterbuch Englisch-Französisch.* Maschinenlesbare Versionen dieser Wörterbücher wurden den Partnern des Projekts für Forschungszwecke durch die Wörterbuchverlage lizensiert. Mit diesen beiden Wörterbüchern deckt der Prototyp die Sprachrichtungen Englisch-Französisch und Deutsch-Englisch ab. Aus lizenzrechtlichen und arbeitsökonomischen Gründen werden jedoch nur Ausschnitte dieser Wörterbücher für den Prototyp benutzt.

Technische Aufbereitung der Wörterbücher

Die Verlage stellten als maschinenlesbare Versionen der Wörterbücher aufbereitete Satzbanddateien zur Verfügung. Um gezielt auf die in den Wörterbuchartikeln enthaltenen Informationen zugreifen zu können, müssen die Artikel vollständig strukturell aufbereitet werden. Dazu wurde der Wörterbuchparser *LexParse* verwendet, der mittels einer benutzerdefinierten Grammatik die hierarchische Mikrostruktur von Wörterbuchartikeln erkennen und explizit darstellen kann. Die für die zwei Wörterbücher entwickelten *LexParse-Grammatiken* decken möglichst umfassend alle Strukturtypen der Wörterbuchartikel ab und schließen inkonsistente und fehlerhafte Artikel aus, die einen beträchtlichen Teil des Wörterbuchs ausmachen. Diese fehlerhaften Artikel wurden manuell korrigiert und erneut geparst. Die resultierenden SGML-annotierten Wörterbücher konnten nun zusammen mit der von LexParse erzeugten DTD in einem SGML-Editor lexikographisch aufbereitet werden.

Teils beim Parsen, teils in einer Nachbearbeitung wurden einige Dekompaktierungen und Markierungskorrekturen vorgenommen. Für die Erstellung eines Index mußten Lemmavarianten ausgeschrieben und Subartikel (Nischeneinträge) aufgelöst werden. Diese Arbeiten wurden größtenteils automatisch durch-

geführt. Abschließend wurden die so entstandenen »Wörterbuchdatenbanken« in eine für beide Wörterbücher gleiche Datenstruktur überführt, die vom Nachschlagesystem *Locolex* benutzt wird.

Lexikographische Erweiterungen

Um aus den geparsten Wörterbüchern echte »comprehension dictionaries« zu machen, waren verschiedene lexikographische Anpassungen nötig. Alle Informationen in einem Eintrag, die für das Verstehen des Wortes unnötig sind, müssen explizit markiert sein, damit sie im COMPASS-System unterdrückt werden können. Hierzu gehört

- die explizite Kennzeichnung alternativer, beinahe synonymer Übersetzungen; z. B. wird die komplexe Übersetzungsangabe 'to switch {or} turn {or} put on' für 'einschalten' in drei einzelne Übersetzungen umgeformt und als solche markiert, was ermöglicht, daß COMP ASS die zweite und dritte Übersetzungsvariante verbirgt.
- die Unterscheidung von Verwendungsbeispielen, die nur für die Sprachproduktion wichtig sind, und semantisch komplexen Mehrwortlexemen, die nur als Ganzes verstanden werden können, mittels verschiedener Markierungen.
- das separate Markieren von Angaben zur Übersetzung von präpositionalen Ergänzungen, die innerhalb des Übersetzungselements erscheinen.

Innerhalb einer Bedeutungsgruppe sollte die allgemeinste Übersetzung zuerst genannt werden, damit COMP ASS diese für eine reduzierte Darstellung des Eintrags auswählen kann. Weitere Dekompaktierungen waren nötig, etwa die Ergänzung von direkten Übersetzungen, die aus Platzgründen nur implizit durch Beispielphrasen angegeben waren. Natürlich mußten auch fehlende Variantenformen, fehlende Bedeutungen, komplett fehlende Stichwörter und Mehrwortlexeme (MWL) ergänzt werden, letzteres auf der Basis von Korpusabfragen und der automatischen Extraktion von MWL-Kandidaten aus Textkorpora.

Formalisierung von Kontext-Mustern

Das COMPASS-System soll erkennen, ob das angefragte Wort in einem bestimmten Kontext vorkommt, in dem eine spezielle Übersetzung passend ist, und diese gegebenenfalls auswählen. Damit dies möglich ist, müssen entsprechende kontextuelle Muster im COMPASS-Wörterbuch ergänzt werden. Hierfür wurde im Projekt ein Finite-State-Formalismus von Rank Xerox benutzt, in dem solche Kontext-Muster als reguläre Ausdrücke kodiert werden. Die Kontext-Formalisierung wurde vorerst auf die Erkennung von MWLs und grammatischen Kollokationen beschränkt.

Die Formalisierung erfolgt in mehreren Schritten. Zuerst wird entschieden, welche Kontexte überhaupt formalisiert werden sollen. MWLs und grammatische Kollokationen werden dann in eine sog. kanonische Grundform gebracht,

die auch lexikalische Varianten enthalten kann. Morphologisch flexible Bestandteile werden als solche gekennzeichnet Auf der Basis dieser kanonischen Form wird automatisch ein regulärer Ausdruck generiert, der z. B. die Wortstellungsvariation im Deutschen bereits erfaßt. Besondere Variationsmöglichkeiten eines MWLs werden anschließend von Hand im regulären Ausdruck ergänzt

Das Nachschlagesystem Locolex

Grundlage des Nachschlagesystems ist das von Rank Xerox entwickelte und patentierte System *Locolex.* Der Locolexkern übernimmt die eigentliche Nachschlagearbeit und lädt auf der Basis einer linguistischen Analyse der Wortumgebung die jeweils relevanten Teile eines Wörterbucheintrags. Für einen schnelleren Zugriff auf die einzelnen Wörterbucheinträge wird ein Index der Stichwörter und deren Varianten benutzt. Die Software von Locolex ist weitgehend systemunabhängig. Sie wurde auf verschiedenen Rechnerarchitekturen entwickelt und portiert

Die Komponenten zur linguistischen Analyse der Ausgangssprachen, die sog. Sprachmodelle, sind nicht direkter Bestandteil des Locolexkems. Die Sprachmodelle werden für die jeweiligen Sprachen getrennt entwickelt und in Form von endlichen Automaten an eine Schnittstelle des Locolexkems angebunden. Zu den wichtigsten Komponenten eines Sprachmodells gehören Algorithmen zur morphologischen Analyse und zur Wortartendisambiguierung. Darüberhinaus enthält das Sprachmodell Definitionen von Makros und Variablen für endliche Automaten, die zur Erkennung von Mehrwortlexemen verwendet werden.

Morphologische Analyse

Die morphologische Analyse übernimmt die Reduktion flektierter Wortformen auf deren Grundform und ermöglicht damit den Zugriff auf Wörterbucheinträge auch von flektierten Wortformen aus (z. B. von der Wortform *gesungen* auf das Stichwort *singen).* Darüberhinaus stellt diese Komponente morphosyntaktische Informationen (Wortart, Kasus, Numerus und Genus) bereit, die in den folgenden Analyseschritten für die Auswahl der passenden Bedeutung genutzt werden.

Wortartendisambiguierung

Werden von der morphologischen Analyse mehrdeutige syntaktische Informationen geliefert (z. B. Artikel oder Verb für die Form *einen* bzw. Substantiv oder Verb für Englisch *plan),* dann wird diese Ambiguität von einer Komponente zur Wortartendisambiguierung aufgelöst. Dabei wird ein als *Hidden-Markov-Modell* bekanntes probabilistisches Verfahren verwendet Diese Komponente ist vor allem für Englisch oder Französisch wichtig, wo viele Inhaltswörter bezüglich ihrer Wortart ambig sind.

Laden der relevanten Teile eines Wörterbucheintrags
Die Ergebnisse der morphologischen Analyse und der Wortartendisambiguierung werden für die Auswahl der für den jeweiligen Kontext relevanten Teile eines Wörterbucheintrags genutzt. Über einen Index wird der gesamte Wörterbuchartikel in den Hauptspeicher geladen. Bei diesem Vorgang wird die jeweilige SGML-Struktur des Wörterbuchartikels in eine weitgehend wörterbuchunabhängige, interne Datenstruktur des Systems abgebildet und der von der Disambiguierung ausgewählte Teil besonders gekennzeichnet.

Erkennung von Mehrwortlexemen
Ist das ausgewählte Wort Bestandteil eines Mehrwortlexems und als solches im Wörterbucheintrag kodiert, wird die Übersetzung des gesamten Mehrwortlexems und nicht des einzelnen Wortes geliefert. Dies ist ein weiterer Schritt zur Auswahl von kontextrelevanten Informationen aus dem Wörterbucheintrag. Hierzu werden die als reguläre Ausdrücke kodierten Mehrwortlexeme des selektierten Wörterbucheintrags mit dem Eingabetext verglichen. Paßt ein regulärer Ausdruck auf den Satzkontext, wird die Übersetzung des zugehörigen Mehrwortlexems speziell gekennzeichnet und zunächst dem Benutzer als Antwort angeboten.

Die graphische Benutzerschnittstelle

Für die Darstellung von Texten und Wörterbucheinträgen wurde eine spezielle graphische Benutzerschnittstelle für Apple-Macintosh-Rechner entwickelt. Kern dieser Benutzerschnittstelle ist der sogenannte *Reader,* ein einfaches Editor-Programm, das es erlaubt, Texte wiederzugeben, einzelne Wörer mit Übersetzungen zu annotieren oder aber auch den Text zu verändern. Entsprechend verfügt dieser Reader über die drei verschiedenen Modi *read, assist* und *edit.*

Für die Anwendung als Lesehilfe ist insbesondere der *assist-Modus* von Interesse. In diesem Modus kann ein Nachschlage- und Analyseprozeß durch einfaches Anwählen eines Wortes mit der Maus aktiviert werden. Als Reaktion auf einen solchen Maus-Klick erscheint ein kleines Hilfsfenster, so in der Nähe des angewählten Wortes plaziert, daß es möglichst wenig Kontext verdeckt. In ihm wird eine Liste derjenigen Übersetzungen aufgeführt, die aufgrund der Kontextanalyse relevant erscheinen.

Dem Benutzer werden im Hilfsfenster verschiedene Optionen angeboten:

- Durch Anwählen einer einzelnen Bedeutungserläuterung wird das Wort im Text mit dieser Erläuterung annotiert Es gibt dabei drei Varianten für die Plazierung der Annotierung, die jeweils vom Benutzer voreingestellt werden können:
  - interlinean der zusätzliche Text erscheint zwischen den Zeilen unter dem erläuterten Wort
  - am Rand: der Text erscheint am Rand auf Höhe der das Wort enthaltenden Zeile

  - separates Fenster: die Bedeutungserläuterungen werden in einem separaten Fenster fortlaufend mitgeschrieben.
- Werden vom Benutzer weitere Informationen zu einer einzelnen Bedeutung gewünscht, kann er sich diese zusätzlichen Informationen durch Anwählen einer für jede Bedeutung vorhandenen Schaltfläche darstellen lassen.
- Schließlich kann durch eine spezielle Schaltfläche der vollständige Wörterbucheintrag ausgegeben werden.
- Unternimmt er nichts, dann bleibt das Hilfsfenster für einen voreingestellten Zeitraum auf dem Bildschirm, bevor es verschwindet.

Die Protokollfunktion

Zusätzlich zur Darstellung der relevanten lexikalischen Informationen auf dem Bildschirm wird eine Reihe von Daten in einer Protokolldatei mitgeschrieben. Dabei können Art und Umfang der zu protokollierenden Daten vom Benutzer bestimmt werden. Diese Funktion erlaubt z. B. ein späteres Rekapitulieren der unbekannten Vokabeln eines Textes.

Benutzertests

Eine erste Evaluierung des Prototyps wurde im Sommer 1995 durch Benutzertests an den Universitäten Bournemouth (Sprachrichtung Deutsch-Englisch) und Lyon 2 (Englisch-Französisch) durchgeführt. Für jede der beiden Ausgangssprachen Deutsch und Englisch standen zwei Zeitungsartikel zur Auswahl, die von Versuchspersonen mit Grundkenntnissen in den Ausgangssprachen mit Hilfe des COMPASS-Systems gelesen wurden. Das Leseverständins der Testpersonenen wurde anschließend durch Verständnisfragen zum Test überprüft. Außerdem wurden die Versuchspersonen mit einem Fragebogen um eine Bewertung verschiedener Funktionen des COMPASS-Systems gebeten.

Erkenntnisse aus der ersten Testphase konnten bei der Entwicklung einer zweiten Version des Prototypen berücksichtigt werden, die wiederum in einer zweiten Testphase im Februar 1996 evaluiert werden konnte. Die Ergebnisse beider Testphasen sind überwiegend positiv ausgefallen.

Projektdaten

Der offizielle Titel des Projekts lautet *COMPASS: Adapting bilingual dictionaries for on-line COMPrehension Assistance.* Das Projekt wurde im Rahmen des Programms *Linguistic Research and Engineering* unter der Nummer 62-080 vom Generaldirektorat XIII der Kommission der Europäischen Gemeinschaft von April 1994 bis April 1996 gefördert Am Projekt waren die folgenden Partner beteiligt:

- Rank Xerox Research Centre, Grenoble (Koordinator)
- Fraunhofer Institut für Arbeit und Organisation, Stuttgart

- Seminar für Sprachwissenschaft, Universität Tübingen
- Department of Marketing, Advertising and Public Relations, Bournemouth University
- Langues Étrangères appliquées, Université Lyon 2