Uta Seewald-Heeg

Vorwort

Sprachtechnologie für die multilinguale Kommunikation hat mittlerweile Eingang in zahlreiche Produkte gefunden, die in vielen Bereichen der Wirtschaft zu einem selbstverständlichen Instrumentarium geworden sind. Gründe hierfür liegen unter anderem im Zusammenrücken der Märkte und dem rasanten Gebrauch des Internet als Recherche-, Informations- und Präsentationsmedium. Die Anforderung an effiziente multilinguale Kommunikation in all ihren Facetten stellt insbesondere auch für die Computerlinguistik eine Herausforderung dar, denn die stets wachsende Menge elektronisch verfügbaren Wissens macht Werkzeuge zur Aufbereitung des Wissens, zum Wissensmanagement und zur maschinellen Sprachverarbeitung notwendig. Computerlinguistische Forschung und Entwicklung ist zu Beginn des 21. Jahrhunderts nicht nur eine akademische Übung, sondern inzwischen auch zu einem Gegenstand wirtschaftlichen Interesses geworden.[1]

Das Spektrum der in diesem Band versammelten Beiträge, die im Rahmen der Frühjahrstagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV)[2] vom 26. bis 28. März 2003 an der Hochschule Anhalt[3] in Köthen präsentiert wurden, umfasst Fragen von sprachtechnologischer Relevanz, die von konkreten Einsatzfeldern wie der Übersetzung bzw. Lokalisierung von Produkten über korpuslinguistische Fragestellungen bis hin zu theoretischen Erörterungen reichen.

Im ersten Teil des Bandes, der unter dem Titel „Young Researchers – Abschlussarbeiten“ steht, werden die Ergebnisse von fünf Abschlussarbeiten präsentiert, die in die Vorauswahl des Wettbewerbs um den GLDV-Preis 2003 für die beste Abschlussarbeit gekommen sind. Die Publikation des Tagungsbandes zur GLDV-Frühjahrstagung enthält eine CD, auf der die Abschlussarbeiten auch jeweils in ihrer als Diplom- oder Magisterarbeit eingereichten Langform enthalten sind, um die in den Arbeiten vorgestellten Innovationen und Verfahren auch im Detail studieren zu können.

Die vorgestellten Arbeiten reichen von Fragestellungen der Wissensextraktion bis hin zur Textgenerierung und der akustischen Präsentation von Webinhalten: Die Wissensextraktion aus elektronisch verfügbaren Korpora ist sowohl Gegenstand der Arbeit von Christian Biemann, der ein Lernverfahren vorstellt, mit dem aus großen Textkorpora semantische Relationen automatisch extrahiert werden, als auch der Arbeit von Stefan Bordag, der sich mit der Auflösung von Ambiguitäten auf lexikalischer Ebene

---

[1] Das belegen auch zahlreiche Beiträge des Sammelbandes „Computerlinguistik – Was geht, was kommt?“, der von Gerd Willée, Bernhard Schröder und Hans-Christian Schmitz als Festschrift für Winfried Lenders Ende 2002 herausgegeben wurde.

[2] http://www.gldv.org

[3] http://www.inf.hs-anhalt.de

beschäftigt und hierzu grafentheoretische Ansätze ausnutzt, um mittels eines Approximationsverfahrens Gebrauchskontexte einzelner Wortformen errechnet. Auch die Arbeit von David Reitter beschäftigt sich mit Fragen der Wissensextraktion. Er entwickelt einen Analysealgorithmus, der die Kohärenzstruktur eines Textes mit Hilfe der Rhetorical Structure Theory analysiert.

In einen dem Text-Mining komplementären Bereich, den Bereich der Textgenerierung, fällt die Arbeit von Christian Chiarcos, der eine Satzplanungskomponente entwickelt, die auf einem kognitiv plausiblen und kohärenten Modell der Sprachproduktion aufbaut, das auf dem Konzept der Salienz als einer Komponente der Topologie mentaler Diskursmodelle beruht. Der akustischen Präsentation von Webinhalten ist schließlich die Arbeit von Tobias Göbel gewidmet. Er entwickelt in seiner Arbeit ein Verfahren, das Web-Seiten automatisch in VoiceXML-Dokumente transformiert, um sie im Auto, am Telefon oder sehbehinderten Menschen verfügbar machen zu können.

Der zweite Teil des vorliegenden Bandes versammelt Beiträge, die unterschiedliche Aspekte der Sprachtechnologie für die multilinguale Kommunikation behandeln und sich im weitesten Sinne den Gebieten Lokalisierung bzw. Übersetzung, Lexikografie und Retrieval zuordnen lassen. Dieser Teil wird eröffnet mit einem Beitrag von Reinhard Schäler, der Faktoren und Entwicklungstendenzen aufzeigt, die den erfolgreichen Einsatz von Sprachtechnologie in dem jungen Wirtschaftszweig der Lokalisierungsindustrie sicherstellt. Die nachfolgenden Beiträge dieses Teils sind zum einen Fragen der maschinellen Lexikografie, der maschinellen bzw. maschinengestützten Übersetzung sowie Retrieval-Problemen bei der Recherche im Internet gewidmet.

Dieter Seelbach beschäftigt sich in seinem Beitrag mit den Spezifika von Partikelverben und Verben mit typischen Adverbialen des Sprachpaars Deutsch–Französisch, um auf dieser Grundlage Einträge für ein bilinguales elektronisches Lexikon zu erarbeiten. Stefan Langer untersucht die Kodierung der Hyponymierelation in elektronischen Wörterbüchern und zeigt Lösungsmöglichkeiten zu deren korrekter lexikografischer Erfassung auf. Nico Weber analysiert in seinem Beitrag Wörterbücher maschineller Übersetzungssysteme und zeigt anhand konkreter Übersetzungsbeispiele Möglichkeiten der Qualitätssteigerung maschineller Übersetzung auf. Andrea Abel und Leonhard Voltmer beschäftigen sich in ihrem Beitrag mit der Evaluierung der italienisch-deutschen Internetschnittstellen zu einem deutsch-italienischen Lernerwörterbuch und einer Datenbank für Rechtsterminologie. Julie Carson-Berndsen und Moritz Neugebauer stellen ausgehend vom Konzept einer *multilingual time map* ein portables, generisches in XML kodiertes Lexikonformat vor, das phonologische Information für beliebige Anwendungen in unterschiedlicher Qualität bereitzustellen vermag.

Kurt Eberle diskutiert in seinem Beitrag ein um einen semantischen Filter erweitertes Resolutionsverfahren zur Anaphernresolution, das einen Kompromiss zwischen semantischer Repräsentation bzw. Auswertung und effizienter Verarbeitung im maschinellen Übersetzungsprozess oder beim Retrieval anstrebt. Felix Sasaki beschäftigt

sich in seinem Beitrag mit Textauszeichnungen im Original und in der Übersetzung und nutzt dabei die Ausdrucksmöglichkeiten von Schemasprachen, die durch ein als CSD (*Context Specification Document*) bezeichnetes Format ergänzt werden.

Die beiden letzten Arbeiten dieses Teil beschäftigen sich mit Eigenschaften von Information-Retrieval-Systemen, die in Evaluierungsprojekten näher untersucht wurden: Thomas Mandl und Christa Womser-Hacker stellen das europäische Evaluierungsprojekt CLEF (*Cross Language Evaluation Forum*) vor und gehen der Frage nach, inwieweit linguistische Eigenschaften der Anfragen Rückschlüsse auf die Qualität der Retrievalergebnisse zulassen. Ursula Fissgus und ihr Autorenteam analysieren im Internet verfügbare Suchmaschinen auf ihre Einsetzbarkeit für die multilinguale Recherche vor dem Hintergrund, dass Suchanfragen mit transliterierten Sonderzeichen erfolgen müssen.

Im dritten Teil des vorliegenden Bandes kommen unter dem Titel „Korpuslinguistik – Texttechnologie“ Fragestellungen zur Sprache, die unterschiedliche Aspekte der Theoriebildung, der Wissensextraktion bzw. des Text-Mining sowie der Textauszeichnung behandeln. Hier finden sich neben Arbeiten, die sich mit theoretischen Problemstellungen beschäftigen, auch solche, die Entwicklungen und Anwendungen vorstellen, die auf korpuslinguistischen Arbeiten aufbauen.

Am Anfang dieses Teils steht der Beitrag von Vladislav Kuboň, der die Struktur des tschechischen Nationalkorpus (CNC) sowie der Prager Dependenzbaumbank (PDT) vorstellt und darlegt, dass Ergebnisse und Werkzeuge der Arbeit an monolingualen Korpora auch für multilinguale Anwendungen, etwa im Bereich der Lokalisierung, einsetzbar sind. Einen stärker theoretischen Akzent setzen die folgenden Arbeiten: Reinhard Köhler entwirft in seinem Beitrag ein Basismodell eines im Rahmen der synergetischen Linguistik aufgestellten syntaktischen Subsystems der Sprache. Alexander Mehler entwirft in seinem Beitrag ein Fundament für korpusanalytische Ansätze in der Semantik. In den Bereich der Semantik fällt auch die Arbeit von Reinhard Rapp, der in seinem Aufsatz ein Verfahren zur maschinellen Bestimmung der Bedeutungen eines mehrdeutigen Wortes durch die Analyse von Unterschieden seiner lexikalischen Umgebungen in Textkorpora verschiedener Textsorten entwirft und hierzu die erst in jüngerer Zeit entwickelte Unabhängigkeitsanalyse (*Independent Component Analysis, ICA*) einsetzt. Arne Ziegler beschäftigt sich in seinem Beitrag mit der linguistischen Textstrukturanalyse und beschreibt ein Modell, mit dem kognitiv relevante Ordnungsmuster in Texten ermittelt werden.

Die folgenden Beiträge des dritten Teils erörtern texttechnologische Fragestellungen, wobei XML als Annotationsformulismus jeweils im Vordergrund steht. Sven Naumann richtet sein Augenmerk auf syntaktische Annotationen, während Daniela Goecke, Daniel Naber und Andreas Witt einen Ansatz zur Auswertung XML-annotierter Dokumente, die hinsichtlich verschiedener linguistischer Ebenen annotiert sind, vorstellen. Ulrike Gut, Jan-Torsten Milde und Karola Pitsch beschreiben in ihrem

Beitrag ein System zur Erstellung und Analyse multilingualer, multimodaler bilingualer Korpora gesprochener Sprache. In diesem Zusammenhang beschreiben sie sowohl die Struktur des XML-Formats als auch die hier zum Einsatz kommenden Software-Komponenten der TASX-Umgebung, eines Systems zur Erstellung und Auswertung von großen Sprachkorpora. Thomas Schmidt stellt in seinem Beitrag ein XML-System zur computergestützten Diskurstranskription für ein mehrsprachiges Diskurskorpus vor, das dazu dient, Formen der Semikommunikation zu untersuchen und das aus Transkriptionen von Radiosendungen, Interviews usw. besteht, in denen Sprecher des Dänischen, Norwegischen oder Schwedischen in ihrer jeweiligen Muttersprache miteinander kommunizieren. Peter Koepke und Bernhard Schröder untersuchen ein Korpus mathematischer Beweise auf die verwendeten linguistischen Mittel hin. Zur Auszeichnung des Korpus nach formalsemantischen Gesichtspunkten wird auf der Grundlage von XML eine Annotationssprache, ProofML, entwickelt.

Die beiden letzten Arbeiten widmen sich schließlich Verfahren zur Wissensextraktion. Uwe Quasthoff, Matthias Richter und Christian Wolff beschreiben in ihrem Beitrag einen Web-Service, der eine Text-Mining-Engine zur Auswertung von Online-Pressetexten nutzt, und Manuela Kunze und Dietmar Rösner stellen eine Dokumentworkbench (XDOC) vor, die Methoden der Computerlinguistik und der Wissensverarbeitung nutzt, um effektives Information Retrieval und Informationsextraktion zu ermöglichen.