Claudia Kunze, Lothar Lemnitzer, Andreas Wagner

Einleitung

Der vorliegende Band enthält alle Beiträge des 1. GermaNet-Anwender-Workshops, der vom 09.–10.10. 2003 in Tübingen stattgefunden hat. Wir blicken auf eine sehr produktive und anregende Veranstaltung mit zahlreichen Diskussionen zurück, Dank des großen Engagements aller Beteiligten. In diesem Vorwort wollen wir kurz die Entwicklung skizzieren, die zur Realisierung unserer Tagung geführt hat.

Seit dem unerwarteten Erfolg, den das Princeton WordNet (vgl. Miller et al. 1990, Fellbaum 1998) als „Mutter aller Netze“ in Computerlinguistik und Sprachtechnologie verbuchen konnte, sind zahlreiche einzelsprachliche und auch multilinguale Wortnetze in nationalen und internationalen Projekten entwickelt worden, z.B. *ItalNet, Dutch WordNet, Portuguese WordNet, EuroWordNet, BalkaNet*, um nur einige zu nennen.

Wortnetze im Stile des Princeton WordNets bilden die häufigsten und wichtigsten Wörter einer Sprache und ihre bedeutungstragenden Beziehungen zu anderen Wörtern der Sprache ab. Ein zentraler Anwendungsbereich ist die Lesartendisambiguierung, welche eine unabdingbare Voraussetzung für Anwendungen im Bereich der Informationserschließung und der Maschinellen Übersetzung, für die semantische Annotierung von Sprachdaten und die Entwicklung verschiedener Werkzeuge zum Sprach- und Informationserwerb darstellt.

In Tübingen hatte Helmut Feldweg die Idee, ein Wortnetz für das Deutsche zu erstellen, und konnte 1995 erste Projektmittel vom Land Baden–Württemberg einwerben. Mit großem Engagement begann das erste LexikographInnen-Team um Helmut - darunter Birgit Hamp, Michael Hipp, Susanne Schüle, Rosmary Stegmann, Christine Thielen, Valérie Béchet-Tsarnou - unter der Anleitung Christiane Fellbaums den Aufbau des GermaNet (vgl. Hamp & Feldweg 1997). Durch den Weggang mehrerer Mitarbeiter, aber vor allem durch Helmuts plötzlichen, viel zu frühen Tod, drohte das GermaNet–Projekt zu scheitern, weil niemand mehr zur Weiterentwicklung und Pflege der Daten verfügbar war.

Im Zuge der noch von Helmut akquirierten Projektbeteiligung am EuroWordNet-2 Vorhaben kamen mit Andreas Wagner und Claudia Kunze zwei neue Mitarbeiter, die ab 1998 die Integration des deutschen Wortnetzes in die multilinguale Datenbank EuroWordNet vornahmen und GermaNet selbst in dieser Zeit signifikant erweitern und verbessern konnten - nicht zuletzt durch die Standardisierungsbemühungen des internationalen Konsortiums unter der Leitung von Piek Vossen (vgl. Vossen 1999, Wagner & Kunze 1999) und den Abgleich mit anderen Wortnetzen.

Nach dieser Phase produktiver Kooperation mit europäischen Partnern wurde die Weiterentwicklung des GermaNet durch Karin Naumann und Claudia Kunze in einem zweijährigen Projekt erneut vom Land Baden–Württemberg gefördert (vgl. Kunze 2001). Auf Initiative von Lothar Lemnitzer entstand eine XML-Version der GermaNet ‚Lexicographers' Files', auf dessen Basis ein Visualisierungstool erstellt wurde (vgl. Kunze & Lemnitzer 2002a, Kunze & Lemnitzer 2002b).

GermaNet ist mittlerweile eine stattliche Ressource mit ca. 42 000 Synsets und mehr als 61 000 Lesarten, die von zahlreichen akade-

mischen und industriellen Anwendern genutzt wird. Uns interessierte zunehmend, welche Erfahrungen die GermaNet-Nutzer mit GermaNet in konkreten Anwendungen machen; dagegen ist das Feedback nach der Lizenzierung der Daten seitens der Anwender oftmals nicht mehr allzu groß.

Auf der Language Ressources and Evaluation Conference (LREC) 2002 wurde im Rahmen eines von uns gemeinsam mit der BalkaNet Gruppe durchgeführten Workshops über „Wordnet structures and standardization" (vgl. Lemnitzer et al. 2002) von mehreren Teilnehmern angeregt, in Tübingen einen GermaNet User Workshop zu veranstalten; eine Idee, die wir gern aufgegriffen haben. Dank einer Zusage des Vorstandes der „Gesellschaft für linguistische Datenverarbeitung" konnten wir unser Vorhaben als GLDV-Workshop realisieren und ankündigen.

Die 14 Beiträge haben wir inhaltlich auf drei Sektionen verteilt, wobei diese jeweilig recht heterogene Themen und Ansätze umfassen.

Die erste Sektion fokussiert Anwendungen des GermaNet und des EuroWordNet in verschiedenen, vorwiegend computerlinguistischen Szenarios.

„Using EWN within the Speech Operated System EMBASSI" von Iman Thabet, Bernd Ludwig, Frank-Peter Schweinberger, Kerstin Bücher und Günter Görz thematisiert inkrementelle semantische Konstruktion auf Grundlage der Description Logics für ein Dialogsystem gesprochener Sprache. In diesem Ansatz fungiert EuroWordNet als lexikalische Hintergrundressource zur Modellierung der linguistischen Daten, welche in verschiedenen Systemkomponenten, internen Domänen und Anwendungen wiederverwendet werden sollen. Die AutorInnen geben einen umfassenden Überblick über die semantische Verarbeitung in ihrem Dialogsystem und die Rolle, die EuroWordNet Synsets darin spielen. Sie formulieren Desiderate an ein ideales Wortnetz als geeignete Wissensbasis.

Manuela Kunze und Dietmar Rösner berichten in ihrem Paper „Issues in Exploiting GermaNet as a Resource in Real Applications" über ihre Erfahrungen mit GermaNet als Hintergrundressource für die domänenspezifische Dokumentenanalyse, die sie anhand einer Testsuite von Autopsieprotokollen exemplifizieren. Sie stellen Überlegungen zur Integration des GermaNet in die XML-basierte Dokumentenverarbeitung (XDOC) an und schließen mit einer Wunschliste an die GermaNet-Entwickler, die sich vor allem um einen verbesserten Zugriff auf orthographische und morphologische Varianten der Suchwörter zentriert.

In seinem Beitrag „Die semantische Auswertung von Produktanforderungen mit Hilfe des GermaNet" skizziert Matthias Jörg, wie GermaNet ein industrielles System („ReMaS") unterstützen kann, das für die Analyse auch natürlichsprachlich vorliegender Produktanforderungen als Ergebnis ein semantisches Anforderungsnetzwerk für Produktentwickler, z.B. Automobilhersteller, erzeugen soll. GermaNet wird bei der Auswertung der Anforderungsmodellierung, die im Paper aus Anwenderperspektive umfassend dargestellt wird, zum Einsatz kommen.

Der Nutzen des GermaNet für andere lexikalisch-semantische Ressourcen wird im Beitrag „Die Verwendung zur Pflege und Erweiterung des Computerlexikons HaGenLex" von Rainer Osswald thematisiert. Mittels Lesartenzuordnung zwischen Einträgen aus GermaNet und HaGenLex wird die Abdeckung beider Ressourcen verglichen und die Aufdeckung von lexikon-internen Inkonsistenzen ermöglicht. Diese Methode gestattet die Übernahme der lexikalisch–semantischen Relationen aus GermaNet in die ebenfalls semantisch basierte HaGenLex-Struktur. Der Aufsatz deckt Schwachstellen in den GermaNet-Hierarchien bezüglich der aspektuellen Charakterisierung von Verben und der Kreuzklassifikation von Partikelverben auf und schlägt in diesen Bereichen eine Restrukturierung vor.

Einleitung

Um kognitiv inspirierte Wissensmodellierung geht es beim Mapa (=Mapping Archtecture for People's Association)- Studienprojekt, das Kathrin Beck in ihrem Beitrag „Ein Vokabeltrainer auf der Grundlage von GermaNet und Mapa" in Bezug auf die im Titel genannte Beispielanwendung vorstellt. Mapa ist eine Plattform, welche die benutzerfreundliche Modellierung von Wissensnetzen ermöglicht, im Falle des hier vorgestellten Vokabeltrainers die GermaNet Daten als Netzstruktur, auf deren Grundlage deutsche Vokabeln „im Kontext" verstanden und gelernt werden können.

Besonders freuen wir uns über den künstlerisch inspirierten Beitrag von Daniela Alina Plewe, Steffen Meschkat und Manfred Stede, die das Projekt „GeneralNews – ein Metabrowser" präsentieren. In Realzeit ersetzt der beschriebene Metabrowser Wörter auf Webseiten durch ihre Synonyme, Hyperonyme oder Hyponyme und erzeugt dadurch ähnliche, abstraktere oder spezifischere Texte, die somit neue Beschreibungen der Welt generieren bzw. die Perspektive auf „neue Welten" eröffnen. Als Hintergrundressource lexikalischen Wissens fungiert WordNet (bzw. GermaNet für eine deutsche Version des Metabrowsers).

Auch die zweite Sektion thematisiert Anwendungen des GermaNet oder WordNet, wobei hier jeweilig die quantitativen Methoden im Vordergrund stehen.

Sabine Schulte im Walde zeigt mit „GermaNet Synsets as Selectional Preferences in Semantic Verb Clustering", wie statistische Verfahren zu einer verfeinerten automatischen Verbklassifizierung führen können, wenn die Argumentpositionen von diathetischen Verben mit Information über ihre selektionalen Präferenzen in Form von semantischen Topknoten aus GermaNet angereichert werden. Sie vergleicht drei verschiedene Ausgangsebenen für die Verbklassifikation (eine rein syntaktisch motivierte Vorklassifikation, eine syntaktische Vorklassifikation, die geforderte und freie Präpositionalphrasen integriert, und schließlich die zusätzliche Ausstattung der Argumentpositionen mit semantischen Knoten, welche die selektionale Präferenz ausdrücken), und deren Eignung für die Erkennung semantischer Verbcluster.

Andreas Wagner diskutiert in seinem Beitrag „Estimating Frequency Counts of Concepts in Multiple-Inheritance Hierarchies" verschiedene Verfahren zur Häufigkeitsschätzung von Konzepten in Wortnetzen mit Hilfe von Korpusdaten. Insbesondere werden Probleme angesprochen, die sich in diesem Zusammenhang durch die multiple Vererbungsstruktur von Wortnetzen ergeben, und es wird ein neuer Lösungsansatz für diese Probleme vorgeschlagen. Da multiple Vererbung (Kreuzklassifikation) ein wesentliches Strukturierungsprinzip in GermaNet darstellt, ist der vorgestellte Ansatz für das deutsche Wortnetz besonders relevant.

In ihrem Beitrag „Domain Specific Sense Disambiguation with Unsupervised Methods" stellen Diana Steffen, Bogdan Sacaleanu und Paul Buitelaar Experimente zur semantischen Disambiguierung von Wörtern in medizinischen Texten vor. Die Lesarten werden aus GermaNet entnommen und in unüberwachten Verfahren disambiguiert. Die Autoren kombinieren verschiedene Disambiguierungsverfahren (vertikal und horizontal) sowie die Auswertung verschiedener Parameter, z.B. Größe und Struktur der Korpora.

Ein Verfahren zum Lernen von Relationen steht im Mittelpunkt des Papers „Lernen paradigmatischer Relationen auf iterierten Kollokationen", das Christian Biemann, Stefan Bordag und Uwe Quasthoff präsentieren. Relationen zwischen Wörtern wie Synonymie, Hyponymie und Kohyponymie werden auf der Basis einer Trainingsmenge, welche die passenden semantischen Merkmale liefert, gelernt, um in mehreren Schritten lexikalisch-semantische Wortnetze (semi-)automatisch zu erweitern. Dabei

wird im Gegensatz zu anderen Verfahren bei der Trainingsmenge nicht nur von einem relevanten Merkmal, sondern von mehreren geeigneten Features ausgegangen.

Die dritte Sektion ist mit strukturbezogenen Aspekten zu Wortnetzen befasst, und zwar in Hinblick auf Repräsentationsvarianten wie TermNet und TopicMaps, Wortnetzverknüpfungen wie zwischen GermaNet und UniNet und die Integration mit anderen lexikalisch-semantischen Ressourcen wie FrameNet.

Maren Runte, Michael Beißwenger und Angelika Storrer thematisieren mit der „Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet" die aus dem Projekt HyTex erwachsene Konzeption eines domänenspezifischen semantischen Wissensnetzes für Texttechnologie und Hypermedia. Die Struktur des vorgestellten TermNet unterscheidet im Gegensatz zu WordNet zwischen lexikalischen und ontologischen Relationen, wodurch das automatische Linking von Wörtern in Fachtexten zu den geeigneten Domänenkonzepten verbessert werden kann.

Der Beitrag „Von der Erstellung bis zur Nutzung: Wortnetze als XML Topic Maps" von Eva Anna Lenz, Benjamin Birkenhage und Jan Frederik Maas zeigt, wie Wortnetze als XML Topic Maps repräsentiert und visualisiert und zum Hypertext-Linking genutzt werden. Da XML Topic Maps als Repräsentationsformat in Bezug zu anderen Repräsentationsvarianten für Wortnetze betrachtet wird, liefert diese Arbeit Überlegungen zum wichtigen Thema der Standardisierung semantischer Netze.

Die Erweiterung von Wortnetzen wie GermaNet steht im Mittelpunkt des Papers „GermaNet und UniNet: Anknüpfen an semantische Netze" von Simon Clematide. Der Autor argumentiert, dass für einen effizienten Einsatz in sprachtechnologischen Anwendungen das allgemeinsprachlich ausgerichtete GermaNet um Domänensprachen erweitert werden muss. In diesem Zusammenhang wird das UniNet der Universität Zürich, das Begriffe aus dem Bereich der Hochschule kodiert, und seine Integration mit GermaNet problematisiert.

Mit Petra Steiners Beitrag „FrameNet und WordNet, Perspektiven für die Verknüpfung zweier lexikalisch–semantischer Netze" wird ein weiterer Ansatz zur Integration von semantischen Ressourcen diskutiert. Die Autorin stellt das FrameNet-Projekt, Strukturen des FrameNet und die Gewinnung lexikalischer Information in diesem Ansatz vor, und präsentiert Überlegungen zur Verknüpfung der komplementären Informationen, die in FrameNet und GermaNet enthalten sind, in einem integrierten Datenbankmodell.

Wir danken der GLDV für ihre großzügige Förderung, durch welche wir diesen Workshop und seine Proceedings ohne Tagungsgebühren anbieten konnten. Dank gebührt Nicole Maruschka, Daniela Stelle und Zoulia Rakhmatoullina, welche bei der Vorbereitung des Workshops unermüdlich geholfen haben; Cornelia Stoll für ihren organisatorischen Beistand und Erhard Hinrichs für seine ideelle Unterstützung. Wir danken allen Teilnehmern, die zum Workshop nach Tübingen gekommen sind und mit ihren Paper- und Diskussionsbeiträgen diese Tagung erst möglich gemacht haben.

Literatur

Fellbaum, Ch. (ed.) (1998). WordNet – An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, MA / London: MIT Press.

Hamp, B.; Feldweg, H. (1997). "GermaNet - a Lexical-Semantic Net for German." In: Vossen, P. et al. (Hrsg.) (1997). Proceedings of the ACL / EACL-97 Workshop on Automatic Information Extraction and Buliding of Lexical-Semantic Resources for NLP Applications, 9-15.

Kunze, C. (2001). „Lexikalisch-semantische Wortnetze." In: Carstensen, K.-U. et al. (Hrsg.) (2001). Computerlinguistik und Sprachtechnologie: Eine Einführung. Heidelberg: Spektrum Akademischer Verlag, 386-393.

Kunze, C.; Lemnitzer, L. (2002a). "GermaNet - Representation, Visualization, Application." In: Proceedings LREC 2002. 3rd International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain, May/June 2002, 1485-1491.

Kunze, C.; Lemnitzer, L. (2002b). "Standardizing Wordnets in a Web-compliant Format: The Case of GermaNet." In: Christodoulakis, Kunze & Lemnitzer (2002), 24-29.

Christodoulakis, D.N.; Kunze, C.; Lemnitzer, L.; (Hrsg.) (2002). Proceedings of the Workshop on Wordnet Structures and Standardizations, and how these Affect Wordnet Applications and Evaluation. 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain, 28th May 2002.

Miller, G. A. et al. (1990). Five Papers on WordNet. Technical Report 43, Princeton University, Cognitive Science Laboratory, *ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf* [accessed April 2004, first published in: Journal of Lexicography 3(4) (1990), 235-312].

Vossen, P. (Hrsg.) (1999). EuroWordNet: A Multilingual Database with Lexical-semantic Networks. Dordrecht: Kluwer Academic Publishers.

Wagner, A.; Kunze, C. (1999). "Integrating GermaNet into EuroWordNet, a Multilingual Lexical-Semantic Database." In: Sprache und Datenverarbeitung, 23(2) (1999), 5-20.