Simon Clematide

GermaNet und UniNet

Abstract
Relationale lexikalische Semantik in der Tradition von WordNet ist Lexikographie, welche im Spannungsfeld von Ontologie und Terminologie arbeitet. Um semantische Netze wie das GermaNet, welche orientiert sind auf den Grundwortschatz, für Anwendungen wie Informationserschliessung nutzbar zu machen, sind Erweiterung und Spezialisierung unabdingbar. Dies versucht das UniNet für das Sachgebiet „Hochschulen und ihre Administration". Die Problematik der Integration von Netzen wird allgemein sowie hinsichtlich des GermaNet/UniNet diskutiert.

1. Einleitung

Ausgehend von WordNet (Fellbaum 1998) ist in den letzten Jahren eine Vielzahl von Projekten entstanden, welche diese Art von Ressource für verschiedenste Einzelsprachen adaptieren oder WordNet direkt übersetzen (siehe *http://www.globalwordnet.org*). Für Deutsch liegt mit GermaNet (Hamp & Feldweg 1997; Kunze 2000) eine Adaption vor, welche den Grundwortschatz abdecken soll und über EuroWordNet multilingual integriert ist.

Für die direkte praktische Anwendung in sprachverarbeitenden Systemen sind diese allgemeinen semantischen Netze mit zwei Problemen behaftet:

1. Zu *hohe Allgemeinheit*: Es gibt zu viele Wortbedeutungen, welche im Anwendungsgebiet gar nicht relevant sind.
2. Zu *niedrige Abdeckung*: Es fehlen viele für das Anwendungsgebiet benötigte Wörter.

Das erste Problem lässt sich beheben, indem irrelevante Wortbedeutungen und im Anwendungsgebiet nicht anwendbare semantische Relationen eliminiert werden. Für WordNet mit dem Anwendungsbereich Aviatik haben dies z.B. (Turcato et al. 2000) stark automatisiert gemacht. Andere Ansätze versuchen die einzelnen Wortbedeutungen – wie in der Terminologiearbeit eher üblich – mit Kodes für das Sachgebiet zu ergänzen, was die Zuordnung der möglichen Lesarten eines Wortes im Kontext deutlich erleichtern kann.

Um das zweite Problem zu lösen, müssen die fehlenden Wörter bestimmt und sinnvoll integriert werden. Dies kann wie in (Buitelaar & Sacaleanu 2002) gezeigt für das Anwendungsgebiet Medizin ausgehend von GermaNet recht erfolgreich automatisch gemacht werden; insbesondere, wenn es um das Ergänzen von Unterbegriffen geht. Für das UniNet, das 1999 entstanden ist, wurde allerdings grösstenteils manuell vorgegangen.

2. UniNet

Seine Entstehung verdankt das UniNet den Bemühungen am Institut für Computerlinguistik Zürich, linguistisch fundierte Informationserschliessung im Sinne von Antwortextraktion bzw. *Passage Retrieval* (Arnold et al. 2001) zu machen. Eine Suchmaschine, welche über relativ kleinen Textmengen und einigermassen homogenem Anwendungsgebiet operiert, sollte sowohl beim Indizieren der Suchtexte wie beim Verarbeiten der Anfrage, synonyme Ausdrücke berücksichtigen und verarbeiten können; zudem sollten insbesondere für die (An-)Frage-Expansion auch einfache Hyperonymie-Relatio-

nen zugänglich sein. Die Erfahrungen aus dem englischsprachigen Antwortextraktionssystem (Dowdall et al. 2002) mit WordNet machen deutlich, dass nur eine anwendungsspezifische Ressource Vorteile einbringt.

2.1 Struktur des UniNet

Das UniNet enthält 21.974 Einträge, die aus einem Wort bestehen. Davon sind 73 Namensbezeichnungen, wobei darunter sowohl Eigennamen wie „Schweiz" als auch Bezeichnungen von Institutionen verstanden werden wie „Stiefel-Zangger-Stiftung" oder „Universitätsrat". Die beiden Wörter „Bildungsaktivität" sowie „Lernaktivität" werden als künstlich im Sinn von GermaNet verwendet. Die restlichen 21.899 Einträge sind normale Substantive – ein recht grosser Teil davon ist allerdings automatisch konstruiert aus Kombinationen von wissenschaftlichen Fachgebieten mit unterschiedlichen Kompositionsgliedern wie in „Philosophieprofessor" und „Philosophiestudentin" oder "Philosophievorlesung".

Von den 1.199 mehrteiligen Ausdrücken sind 494 komplexe Namensbezeichnungen wie etwa „schweizerische Eidgenossenschaft" oder „Maturitätsschule für Erwachsene". 12 mehrteilige Ausdrücke sind künstliche Lexikoneinträge wie etwa „sprechende Person" oder „erziehungstätige Person". Die restlichen 693 Einträge sind Nominalgruppen, die teilweise lexikalisierten Gehalt haben wie „zoologischer Garten", teilweise müssten sie eigentlich besser als künstliche Einträge gehandelt werden wie etwa „staatliche Institution".

Im Netz sind je 22.514 Unter- bzw. Oberbegriffsbeziehungen kodiert. Auf der obersten Ebene gibt es 10 Kategorien: „Entität", „Gruppe", „Handlung/Akt", „Ereignis", „Zustand", „Abstraktion", „Besitz", „Phänomen", „Lage", „Ort" sowie „Aspekt des Geistigen", welches „Kognition" und „Motivation" zusammenfasst. Wie in GermaNet kann eine Synonymklasse (*Synset*) zu mehr als einer Synonymklasse in der Oberbegriffsbeziehung stehen. Im UniNet wird davon sogar recht grosszügig Gebrauch gemacht, insgesamt gibt es 6.996 solcher Klassen. Die restlichen 5.379 der insgesamt 12.385 Synonymklassen sind direkt nur mit einem Oberbegriff verbunden. Die 2.648 Meronymie- bzw. Holonymiebeziehungen machen wie im GermaNet keine feinere Unterscheidung in Teil-Ganzes, Element-Menge oder Material-Objekt. Es existieren insgesamt 26 lexikalische Antonymie-Beziehungen.

2.2 Bildungsmuster für Wörter und Synonymklassen

Ein wichtiger Kernbereich des UniNet sind die gut 550 wissenschaftlichen Studienfächer und Disziplinen, welche in einer Taxonomie abgelegt sind. So ist beispielsweise „Rechtswissenschaft" in die Teilfächer „Privatrecht", „Staatsrecht" usw. aufgeteilt, welche wiederum untergliedert sind. Diese Fachbezeichnungen werden automatisch mit Zweitgliedern wie „-studium", „-professorin", „-professor", „-seminar" oder Erstgliedern wie „Hauptfach" verschmolzen und in die entsprechenden Synonymklassen mit den geeigneten semantischen Relationen eingefügt.

Mit dem expliziten Einbau aller Kombinationen bläht man das Netz allerdings mit vielen Wortformen auf, welche kaum je verwendet werden oder als morphologische Unfälle taxiert werden müssen wie etwa die „Hauptfachkatastrophenmedizinerin". Es wäre gerade für semantisch transparente und produktive Wortbildungen sowie für die Verwaltung von Netzen sinnvoll, solche Regularitäten als semantisches Regelwissen ablegen zu können. Allerdings würde man damit den extensionalen Charakter der traditionellen semantischen Netze sprengen.

3 Abdeckung und Lesarten

In diesem Abschnitt werden zwei kleine Experimente beschrieben, welche Umfang und Unterschiede von GermaNet und UniNet in der Anwendung zeigen sollen. Da sich die bei der Konstruktion von UniNet verwendete GermaNet-

Version von 1999 mancherorts recht deutlich unterscheidet von der Version 4 aus dem Jahr 2001, habe ich die Resultate teilweise für beide angegeben.

Grundlage für die Evaluationen war das kleine Korpus „LUIS-Texte", welches von Web-Seiten und elektronischen Broschüren aus dem Umfeld der Hochschulen auf dem Platz Zürich ursprünglich für unser *Passage-Retrieval-System* LUIS zusammengestellt wurde. Daraus sind etwa 350 Sätze mit 6.711 Token ausgewählt, getaggt (STTS-Tagset) und syntaktisch annotiert, welche letztlich noch wortsinndesambiguiert werden sollen.

3.1 Häufige Substantive

In Tabelle 1 sind die häufigsten Substantive aus dem Korpus mit der Anzahl der Synonymklassen abgebildet. Das Zeichen § wird als Wort „Paragraph" behandelt – es ist das einzige Wort, das nicht in das Kerngebiet von UniNet fällt. In 7 Fällen deckt GermaNet 1999 schlechter ab als die neuere Version. Beim Wort „Universität" darf man sich jedoch nicht täuschen lassen. Obwohl GermaNet 2001 und UniNet nur eine einzige Synonymklasse aufweisen, sind dank Mehrfachvererbung letztlich sowohl die Instituts- wie auch die Gebäudelesart präsent. Mit der Einführung mehrerer Oberbegriffsbeziehungen für eine Synonymklasse ist die Gleichung „1 Synset = 1 Lesart" aufgehoben worden.

Obwohl ein Eintrag für „Studierender" fehlt – dieser Ausdruck ist sicher genügend lexikalisiert, um eingetragen zu werden – schneidet GermaNet 2001 im Bereich der häufigsten Substantive im Vergleich zum anwendungsspezifischen UniNet sehr gut ab. Wenn man etwas genauer in die einzelnen Synonymklassen hinein schaut, treten aber einige Unterschiede auf:

So kennt GermaNet 2001 für „Fakultät" eine aus zwei Lesarten von „Fachbereich" als Synonym, UniNet hingegen nicht – dafür vermerkt es noch „Universitätsfakultät" als synonym. UniNet steckt „Prüfung", „Examen", „Test" in eine Klasse, GermaNet 2001 nimmt „Prüfung", „Überprüfung", „Kontrolle" in einer der beiden Lesarten zusammen.

Freq.	Wortform	GN 01	GN 99	UN
38	Fakultät	1	0	1
19	Prüfungen	2	1	2
18	Universität	1	2	1
16	Prüfung	2	1	2
15	§	2	0	0
14	Diplomarbeit	1	0	1
13	Universitäten	1	2	1
12	Studierende	0	0	1
12	Kandidatin	1	1	1
11	Studium	2	0	1
11	Studierenden	0	0	1
11	Latein	1	0	2

Tabelle 1: Anzahl Synonymklassen der häufigsten Substantive im Test-Korpus

3.2 Zufällig ausgewählte Substantive

In einem weiteren Experiment wurden 100 verschiedene Substantive zufällig aus dem Korpus ausgewählt und nach folgenden Kriterien beurteilt:

A Mit wievielen Synsets kommt das Wort in GermaNet 2001 bzw. UniNet vor?
B Ist die relevante Lesart darunter?
C Ist dies eindeutig oder etwas unsicher?
D Ist das Wort Teil eines Mehrwortausdrucks?

In 36 Fällen sind die Informationen von UniNet und GermaNet übereinstimmend. Bei GermaNet fehlen 27, bei UniNet 32 Wörter. In 21 Fällen gibt es das Wort nur im UniNet nicht, aber nur in 3 davon handelt es sich um Ausdrücke, die im Anwendungsgebiet relevant sind. In 16 Fällen gibt es das Wort im UniNet, aber nicht im GermaNet. Dabei sind alle Ausdrücke ausser dem Wort „Ratsuchender" relevant für das Anwendungsgebiet. Für die Wörter „Hinblick", „Literatur" und „Botschaft" kennt UniNet zwar eine Lesart, aber leider die falsche. Dies passiert

GermaNet in 2 Fällen. Im UniNet ist das Urteil, ob eine Lesart wirklich relevant ist, in 6 Fällen unsicher, im GermaNet in 13 Fällen. UniNet liefert 82 Synonymklassen für 68 Wörter (Ambiguitätsrate 1,2), GermaNet 126 für 73 (Ambiguitätsrate 1,7).

4 Verknüpfen von semantischen Netzen

Der urspünglich rein strukturelle Ansatz der relationalen lexikalischen Semantik definiert die Bedeutung eines Wortes nur aus der Menge der semantischen Beziehungen zu den andern Wörtern. Die 1989 erfolgte Einführung definitorischer Glossen im WordNet[1] sowie die Einführung von Beispielsätzen war ein wichtiger Schritt hin bzw. zurück zur traditionelleren Lexikographie.

Die „Selbstorganisation" der Lesarten in semantischen Netzen macht ihre Integration nichttrivial – das Problem ist in unterschiedlicher Form aufgetaucht:

1. **Multilinguale Netze:** Das Verbinden von gleichbedeutenden Wörtern über mehrere europäische Sprachen hinweg war ein Ziel von EuroWordNet (Vossen et al. 1999a). Zu diesem Zweck wurde zuerst die Version 1.5 von WordNet als Referenznetz genommen und einzelsprachliche Synonymklassen mit den entsprechenden Synonymklasse des Referenznetzes indiziert (*Inter-Lingual-Index*). Verschiedene Revisionen des ILI versuchten dann, die Abdeckung und Granularität der Kernbedeutungen zu normalisieren, um ein universell anwendbares Begriffsgerüst zu erhalten.
2. **Teil-Netze**: Die Integration von themen- und anwendungsspezifischen Netzen in allgemeinere Netze ist für viele Anwendungen wichtig. Eine manuelle Integration ist nicht besonders problematisch, wenn die Lexik des anwendungsspezifischen Netzes auf einige wenige Stellen im allgemeineren Netz konzentriert ist – oder falls das allgemeinere Netz sowieso als Grundgerüst verwendet wird. (Magnini & Speranza 2002) enthält ein halbautomatisches Verfahren, um ein bereits bestehendes semantisches Netz aus dem Wirtschaftsbereich ins EuroWordNet einzuhängen.
3. **Netz-Versionen**: Der Ausbau und die Verfeinerung von semantischen Netzen ergeben verschiedene Instanzen, welche sich nicht bloss im Umfang, sondern oft auch in der internen Strukturierung mehr oder weniger stark unterscheiden. Ressourcen, welche auf Version *X* eines semantischen Netzes aufsetzen, lassen sich oft nicht einfach mit Version *Y* koppeln. Ein automatisches Abgleichen unterschiedlicher Versionen ist äusserst wünschenswert – ein erfolgreiches Verfahren für die Abbildung von WordNet 1.5 auf 1.6 stellen (Daudé et al. 2001) vor.

Selbstverständlich können obige Probleme beim Integrieren von Netzen kombiniert auftreten. Ein gutes Beispiel dafür, wie aufwändig dies für verschiedene Revisionen eines multilingualen Netzes ist, findet sich im Kontext von EuroWordNet in (Vossen et al. 1999b).

Wie steht nun das UniNet diesen Problemen gegenüber? Bei der Konstruktion war die Einbettung in das bestehende allgemeinere GermaNet (Version 1999) äusserst nützlich, d.h. ontologische Organisation, viele Überbegriffe und einzelsprachliche Gegebenheiten konnten übernommen werden. Die lexikographische Arbeit liess sich so von einer Person mit vernünftigem Aufwand erledigen. Die Integration von UniNet ins GermaNet war allerdings von Anfang an nie vollständig, d.h. UniNet kann nicht als Teilnetz von GermaNet aufgefasst werden, das als Modul per Knopfdruck die relational strukturierte Terminologie eines bestimmten Sachgebiets zur Verfügung stellt.

Mit der Weiterentwicklung von GermaNet zur Version 4 ist die Kompatibilität zum Uni-

Net vermutlich noch weiter geschrumpft. Wie stark die Abweichungen sind, soll bei uns in einem Studienprojekt abgeklärt werden.

Ein Beispiel: GermaNet 1999 kannte das Wort „Entscheid" oder „Beschluss" nicht. Im UniNet wurden sie dann als Synonym zu „Entscheidung, Festlegung, Bestimmung, Festsetzung" genommen und unter „Akt" subsumiert. In GermaNet 4 kommt sowohl „Beschluss" wie „Entscheid" vor, allerdings in leicht anderen semantischen Beziehungen: Das Wort „Entscheid" wird synonym mit „Bescheid" gesetzt und als Auskunft im Sinn von Information verstanden, die etwas als „Steuerbescheid" oder „Rentenbescheid" auftaucht. Das Wort „Beschluss" dagegen ist bei Kognition als eine Art „Entscheidung, Entschluss" aufgelistet.

Aus heutiger Sicht wäre eine Indizierung der wichtigsten Kategorien mit einem möglichst allgemeinen und stabilen Bedeutungsindex wie etwa dem ILI wünschenswert, weil dadurch eine algorithmische Verknüpfung des UniNet erheblich erleichtert werden könnte.

Ein anderer Punkt ist eine mögliche Erweiterung von UniNet: Durch den Verwendungszweck und den Entstehungsort sind viele Ausdrücke vom schweizerischen Hochschulsystem und Sprachgebrauch geprägt. Es wäre deshalb notwendig, vermehrt regionalsprachliche stilistische Markierungen in den lexikographischen Dateien zu kodieren.

Dank

Geht an Martin Volk für die UniNet-Projektleitung, an Arnold H. Bucher für die Terminologiearbeit und das Kodieren der lexikographischen Dateien, sowie an die bereitwillige Unterstützung des GermaNet-Teams.

Anmerkung

[1] George Miller schreibt im Vorwort zu Fellbaum 1998, dass mit wachsender Grösse von WordNet das Erfassen von Bedeutungen aus der Synonymie-Beziehung allein immer schwieriger wurde und konzediert selbstkritisch: "... definition by synonymy is not adequate."

Literatur

Arnold, T. et al. (2001). „LUIS – ein natürlichsprachliches, universitäres Informationssystem." In: Appelrath, H.-J. et al. (Hrsg.) (2001). Unternehmen Hochschule (UH-01). Bonn: Köllen Verlag [= GI-Edition - Lecture Notes in Informatics (LNI), P-6], 115-126.

Buitelaar, P.; Sacaleanu, B. (2002). "Extending Synsets with Medical Terms." In: Proceedings of the 1st International WordNet Conference, Mysore, India, January 2002.

Daudé, J.; Padró, L.; Rigau, G. (2001). "A Complete wn1.5 to wn1.6 Mapping." In: Proceedings of NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburgh, PA, June 2001.

Dowdall, J. et al. (2002). "Technical Terminology as a Critical Resource." In: Proceedings 3rd International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas de Gran Canaria, Spain, May/June 2002.

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. (2000). "Extension and Use of GermaNet, a Lexical-Semantic Database." In: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, May 2000, 999-1002.

Magnini, B.; Speranza, M. (2002). "Merging Global and Specialized Linguistic Ontologies." In: Proceedings of OntoLex 2002. Workshop held in conjunction with the 3rd International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas de Gran Canaria, Spain, May 2002.

Turcato, D. et al. (2000). "Adapting a Synonym Database to Specific Domains." In: Proceedings of the ACL 2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, Hong Kong, October 2000.

Vossen, P.; Peters, W.; Gonzalo, J. (1999). „Towards a Universal Index of Meaning." In: Standardizing Lexical Resources. Proceedings of the ACL-99 / SIGLEX Workshop, College Park, MD, June 1999.

Vossen, P. et al. (1999). Extending the Inter-Lingual-Index with new concepts. EuroWordNet, Document Nr. LE-4003-2D010, *http://www.illc.uva.nl/EuroWordNet/docs/2D010RTF.zip* [Zugriff April 2004].