GermaNet-Workshop

Do., 19.2.2015, ab 12 Uhr, SFB 991, Universität Düsseldorf

Verena Henrich & Lars Horber
(Seminar für Sprachwissenschaft, Universität Tübingen)

GermaNet ist ein lexikalisch-semantisches Wortnetz der deutschen Sprache, das seit 1997 im Rahmen verschiedener Projekte des Seminars für Sprachwissenschaft an der Universität Tübingen entwickelt wird. Die Modellierung von GermaNet orientiert sich im Wesentlichen an den Strukturierungsprinzipien des englischen Princeton WordNet, wobei die bedeutungstragenden Kategorien Adjektive, Nomina und Verben in semantische Felder subklassifiziert sind. Die zentrale Repräsentationseinheit ist das Synset, in dem lexikalische Einheiten, die das gleiche Konzept ausdrücken, als Menge von Synonymen zusammengefasst werden. Zwischen den Konzepten sind semantische Relationen definiert, die auf das gesamte Konzept oder auf einzelne lexikalische Einheiten Bezug nehmen können.

GermaNet wird als Grundlagenressource für Anwendungen der maschinellen Sprachverarbeitung entwickelt. Die Konzeption von GermaNet basiert auf einer weitgehenden Lesartendisambiguierung, die eine wichtige Voraussetzung für verschiedene computerlinguistische Anwendungen wie maschinelle Übersetzung, Informationserschließung oder die Erstellung von Sprachtechnologietools bildet.

Im Rahmen dieses Workshops wird in drei separaten Vortragsblöcken auf verschiedene Aspekte von GermaNet eingegangen. Die zeitliche Dauer der einzelnen Teile richtet sich flexibel nach den Interessen und Rückfragen der Workshopteilnehmer.

Teil 1: Einführung in GermaNet und Überblick aktueller Arbeiten

Im ersten Workshop-Abschnitt wird eine generelle überblicksartige Einführung in GermaNet gegeben. Die Idee hinter der Modellierung von Synsets und lexikalischen Einheiten sowie die unterschiedlichen Arten von Relationen werden anhand von Beispielen erläutert. Es werden alle vorhandenen Informationen in GermaNet angesprochen, unter anderem die Annotation von Nominalkomposita sowie die Subkategorisierungsrahmen von Verben. Die Workshopteilnehmer erhalten die Möglichkeit sich interaktiv Einträge in GermaNet anzuschauen, um so einen besseren Eindruck von der Ressource zu bekommen.

GermaNet wurde in den vergangenen Jahren auf vielfältige Weise erweitert und in computerlinguistische Anwendungen integriert. Beispielsweise wurden GermaNet-Lesarten (im Rahmen des EuroWordNet-Projekts) manuell mit den Lesarten des englischen Wortnetzes verknüpft; GermaNet-Lesarten wurden semi-automatisch mit Einträgen des Online-Wörterbuchs Wiktionary verlinkt und somit Paraphrasen für GermaNet-Lesarten gewonnen; und Nominalkomposita wurden semi-automatisch in ihre Bestandteile gesplittet und mit linguistischen Eigenschaften annotiert. Außerdem wurde eine Anwendung entwickelt, mit der die semantische Nähe zwischen zwei Begriffen in GermaNet berechnet werden kann. Des Weiteren wurden zwei mit Lesarten annotierte Korpora (ein manuell annotiertes sowie ein semi-automatisch annotiertes Korpus) erstellt und umfassende Experimente mit automatischer Lesartendisambiguierung durchgeführt.

Je nach Interesse der Teilnehmer wird in dieser Einführung auf ausgewählte oder alle genannten Bestandteile und Anwendungen eingegangen.

Teil 2: Linguistische Klassifikation in GermaNet

Im zweiten Teil des Workshops stehen weitergehende linguistische Aspekte – insbesondere zu den Subklassifizierungen der jeweiligen Wortarten – im Vordergrund. Es wird ein allgemeiner Überblick über die semantischen Felder der Adjektive, Nomina und Verben in GermaNet gegeben. Der Schwerpunkt wird auf den Konzeptionsgrundlagen der Adjektivmodellierung liegen. Im Unterschied zur Adjektivklassifikation im Princeton WordNet sind die Adjektive in GermaNet hierarchisch angeordnet, die Vorgehensweise bei der Antonymierelation unterscheidet sich und die Adjektive sind semantischen Feldern zugeordnet. Jedes semantische Feld gliedert sich in weitere Subklassen. So wird zum Beispiel die Klasse der wahrnehmungsspezifischen Adjektive dahingehend subklassifiziert, inwieweit Adjektive Geräusche (geräuschspezifisch), Gerüche (geruchsspezifisch), Beschaffenheiten von Oberflächen (oberflächenspezifisch) oder weitere Eigenschaften kennzeichnen. Die zugrundeliegenden Entscheidungskriterien für die Zuordnung einer Adjektivlesart zu einer bestimmten Klasse bzw. Subklasse werden anhand ausgewählter Beispiele erläutert. Da insbesondere bei den Adjektiven künstliche Konzepte sowie Kreuzklassifikation eine wesentliche Rolle spielen, wird ihre Verwendung ebenfalls thematisiert.

Teil 3: Programmatischer Zugriff auf GermaNet

In dem XML-Format, in dem GermaNet verfügbar gemacht wird, gibt es vier verschiedene Arten von XML-Dateien, die jeweils unterschiedliche Informationen beinhalten: (i) Synsets und lexikalische Einheiten, (ii) konzeptuelle und lexikalische Relationen, (iii) Verknüpfungen zum englischen Princeton Wortnetz, sowie (iv) Paraphrasen aus Wiktionary. Um all diese Inhalte der Ressource in eigenen computerlinguistischen Anwendungen verwenden zu können, ist zumeist ein programmatischer Zugriff erforderlich. Hierfür steht eine Java-API („Application Programming Interface“) zur Verfügung, mit der auf alle Informationen aus GermaNet zugegriffen werden kann.

Zusätzlich gibt es eine zweite Java-API, mit der semantische Nähe in der GermaNet-Graphenhierarchie berechnet werden kann. Diese API implementiert mehrere häufig verwendete Methoden zur Berechnung semantischer Nähe, auf die je nach Interesse eingegangen werden kann.

In diesem dritten Workshop-Teil wird eine Einführung in die Funktionalitäten der beiden APIs gegeben und dann entweder gemeinsam ein GermaNet-Programm entwickelt oder gezielt auf die Fragen und Anforderungen der Workshopteilnehmer eingegangen.

Über die Vortragenden

Beide Vortragenden arbeiten am Lehrstuhl für Allgemeine Sprachwissenschaft und Computerlinguistik (Lehrstuhlinhaber: Prof. Dr. Erhard Hinrichs) des Seminars für Sprachwissenschaft der Universität Tübingen.

Verena Henrich hat einen M.Sc.-Abschluss in Informatik von der Hochschule Darmstadt. Seit 2009 ist sie wissenschaftliche Assistentin und Doktorandin an der Universität Tübingen und unter anderem für die technische Betreuung von GermaNet zuständig. Sie ist federführend verantwortlich für mehrere semi-automatische Erweiterungen der GermaNet-Ressource, wie bspw. das Splitten von Nominalkomposita in ihre Bestandteile oder die Anreicherung von GermaNet-Lesarten mit Paraphrasen aus Wiktionary. In ihrer Doktorarbeit beschäftigt sie sich mit der automatisierten Disambiguierung von GermaNet-Lesarten.

Lars Horber studiert Computerlinguistik an der Universität Tübingen. Als studentische Hilfskraft arbeitet er seit einigen Jahren im lexikographischen Team von GermaNet mit. Er ist vorrangig für Neueinträge von Adjektiven und Nomen und deren Überprüfung zuständig.