IASK: Computerlinguistik

Zusätzliche Lehrveranstaltungen und weitere Informationen werden durch Aushang am schwarzen Brett in INF 325, 1. Stock und auf unseren Internet-Seiten bekanntgegeben.

Fachstudienberater: *Sascha Fendrich* , Do 16.00-17.00 Uhr, Zi. 108, Tel. 06221-543483

Übergreifende Fachkompetenzen

Programmieren I - PI, B02

V/Ü; Nr.: 09-160-04-01; SWS: 4

Di; wöch; 14:15 - 15:45; ab 20.10.2009; INF 306 / SR 13; Hartung, M.

Do; wöch; 16:15 - 17:45; ab 22.10.2009; INF 328 / SR 16; Hartung, M.

Kommentar

Leistungsbewertung:

P1 (Bachelor, neue Prüfungsordnung): 6 LP

B02 (Bachelor, alte Prüfungsordnung): 6 LP

Übergreifende Kompetenzen: 3 LP

Inhalt

Ziel dieser Vorlesung ist, Studierenden einen ersten Überblick über die systematische Entwicklung von wartbaren und korrekten Programmen zu geben. Dies geschieht anhand der objektorientieren, interpretierten Sprache Python, die mit einem einfachen Objektmodell, guter Unterstützung der Modularisierung und einer reichen Bibliothek einen raschen Zugang zu modernen Programmiertechniken und zudem weitgehende Plattformunabhängigkeit bietet. Dabei wird versucht, den Stoff möglichst anhand konkreter (computerlinguistischer) Fragestellungen zu entwickeln.

Themen:

* Programmierung als Problemlösen
* Werte, Typen, Variablen
* Funktionen
* Kontrollstrukturen
* Sequenzen
* Dictionaries
* Klassen und Objekte
* Ausblick auf funktionales Programmieren
* Locales
* Reguläre Ausdrücke
* XML-Behandlung in Python

Leistungsnachweis

* 60% der Übungsaufgaben **müssen** erfolgreich bearbeitet werden
* Abschlussklausur
* Die erfolgreich bestandene Klausur ist Teil der Orientierungsprüfung

Korpuslinguistik - CS-CL, A12

V; Nr.: 09-160-10-08; SWS: 2

Fr; wöch; 11:15 - 12:45; ab 23.10.2009; INF 325 / SR 24; Zielinski, A.

Kommentar

Leistungsbewertung:

CS-CL, BS-CL, BS-AC (Bachelor, neue Prüfungsordnung): 4 LP (Klausur) oder 6 LP (Klausur und Referat)

A12 (Bachelor, alte Prüfungsordnung): 4 LP

Übergreifende Kompetenzen: 2 LP

Inhalt	In der Korpuslinguistik werden linguistische Datensammlungen (Sprachkorpora) systematisch gesammelt und gepflegt, da sie die Basis für linguistische Forschung bilden und zur Überprüfung linguistischer Theorien dienen können. Der Begriff 'Korpus' ist definiert als "eine Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehrerer Sprachen. [...] Die Bestandteile des Korpus, die Texte oder Äußerungsfolgen, bestehen aus den Daten selbst sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind." (Lemnitzer/Zinsmeister). In der Vorlesung geht es um den Einsatz von Korpora in unterschiedlichen Bereichen der Sprachwissenschaft. Ausgehend von den theoretischen Fragestellungen (z. B. in der computerunterstützten Lexikographie oder der Maschinellen Übersetzung) werden grundlegende korpuslinguistische Methoden vorgestellt. Dazu gehören insbesondere effiziente Technologien für die Korpussuche mit Tools wie XAIRA, Cosmas oder TigerSearch als auch Werkzeuge zur quantitativen Analyse (Kookkurrenzanalyse, Translation Memories, etc.).
Leistungsnachweis	Leistungsnachweis ist eine Klausur (4 LP) oder Referat und Klausur (6 LP)
Voraussetzung	Die Teilnehmerzahl für diese Veranstaltung ist begrenzt. Bei zu vielen Teilnehmern haben Studierende der Computerlinguistik Vorrang.
Literatur	* L. Lemnitzer/H. Zinsmeister, Korpuslinguistik: Eine Einführung, Narr, Tübingen 2006 * Ausgewählte Artikel aus: Anke Lüdeling & Merja Kytö (Hgg.) (erscheint 2008): Corpus Linguistics. An International Handbook. Mouton de Gruyter, Berlin. * K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde and H. Langer (ed.): Computerlinguistik und Sprachtechnologie - Eine Einführung. Heidelberg, Spektrum-Verlag. 2001

Vorbereitungskurse

Einführung in die Nutzung computerlinguistischer Ressourcen

Ü; Nr.: 09-160-00-02; SWS: 2

Block; 10:00 - 13:00; 12.10.2009 - 16.10.2009; INF 325 / PCPool; Reiter, N.

Block; 14:00 - 17:00; 12.10.2009 - 16.10.2009; INF 325 / PCPool; Reiter, N.

Kommentar	* begrenzte Teilnehmerzahl * ggf. Vorzug für TeilnehmerInnen am Software-Projekt.
Inhalt	Der Vorkurs vermittelt Grundlagen der Nutzung von Linux-basierten computerlinguistischen Tools und Korpora. Dabei geht es sowohl um allgemeine Linux-Grundlagen (wie z.B. Ein-/Ausgabeumleitung oder nützliche Tools der Linux-Kommandozeile) als auch um einzelne Parser, Tagger, Chunker und andere Hilfstools der Computerlinguistik. Wir werden uns anschauen, wie bestimmte Tools zu benutzen sind, was man aus ihnen herausbekommt (und was nicht) und wie man solche Ausgaben automatisch weiterverarbeiten kann (und zum Beispiel an das nächste Tool weiterverfüttert). Der Kurs beinhaltet Übungen - Wenn es nicht genug Arbeitsplätze für alle gibt, werden TeilnehmerInnen am Softwareprojekt vorgezogen.
Leistungsnachweis	Ungeprüft, unbenoteter Schein
Voraussetzung	Programmierprüfung

Bachelor (alte Prüfungsordnung)

Einführung in die Computerlinguistik - ICL, B01

V/Ü; Nr.: 09-160-01-01; SWS: 4; LP: 6

Di; wöch; 09:15 - 10:45; ab 13.10.2009; INF 350 / OMZ R U013; Frank, A.

Do; wöch; 11:15 - 12:45; ab 22.10.2009; INF 350 / OMZ R U013; Frank, A.

Kommentar

Leistungsbewertung:

ICL (Bachelor, neue Prüfungsordnung): 6 LP

B01 (Bachelor, alte Prüfungsordnung): 6 LP

Inhalt

Die Vorlesung führt ein in die Grundlagen, zentralen Fragestellungen und Methoden der Computerlinguistik. In einem Gesamtüberblick werden die wesentlichen Grundlagen der Computerlinguistik eingeführt:

* Ebenen der Sprachbeschreibung (Phonologie, Morphologie, Syntax, Semantik, Pragmatik),
* formale mathematische und logische Modelle zur Beschreibung der entsprechenden linguistischen Phänomene und
* algorithmische Verfahren zur automatischen Verarbeitung auf Basis dieser Modelle.

Dabei nähern wir uns speziellen Problemen und Fragestellungen der Computerlinguistik und ihren spezifischen Lösungsstrategien. Spezielle Themen werden sein: Ambiguitätsbehandlung, Approximierung sprachlicher Regularitäten, syntaktische und semantische Verarbeitung.

Die Vorlesung gibt einen Überblick über computerlinguistische Anwendungen, diskutiert das Verhältnis zu Nachbardisziplinen, und führt durch praktische Übungen in die speziellen Fragestellungen einzelner Teilgebiete der Computerlinguistik ein.

Leistungsnachweis

* Erfolgreiche Bearbeitung der Übungsaufgaben (mind. 60%)
* Erfolgreich bestandene Klausur
* Aktive Teilnahme

Regelmäßige Präsenz ist Voraussetzung für den Scheinerwerb.

Die erfolgreich bestandene Klausur ist Teil der Orientierungsprüfung.

Literatur

* Daniel Jurafsky and James H. Martin (2000): Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall Series in Artificial Intelligence. Prentice Hall.
* Kai-Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde, Hagen Langer (Hrsg.) (2004): Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg: Spektrum, Akademischer Verlag.
* Natural Language Toolkit, NLTK: http://nltk.sourceforge.net/index.php/Book

Programmieren I - PI, B02

V/Ü; Nr.: 09-160-04-01; SWS: 4

Di; wöch; 14:15 - 15:45; ab 20.10.2009; INF 306 / SR 13; Hartung, M.

Do; wöch; 16:15 - 17:45; ab 22.10.2009; INF 328 / SR 16; Hartung, M.

Kommentar

Leistungsbewertung:

P1 (Bachelor, neue Prüfungsordnung): 6 LP

B02 (Bachelor, alte Prüfungsordnung): 6 LP

Übergreifende Kompetenzen: 3 LP

Inhalt

Ziel dieser Vorlesung ist, Studierenden einen ersten Überblick über die systematische Entwicklung von wartbaren und korrekten Programmen zu geben. Dies geschieht anhand der objektorientieren, interpretierten Sprache Python, die mit einem einfachen Objektmodell, guter Unterstützung der Modularisierung und einer reichen Bibliothek einen raschen Zugang zu modernen Programmiertechniken und zudem weitgehende Plattformunabhängigkeit bietet. Dabei wird versucht, den Stoff möglichst anhand konkreter (computerlinguistischer) Fragestellungen zu entwickeln.

Themen:

* Programmierung als Problemlösen

* Werte, Typen, Variablen
* Funktionen
* Kontrollstrukturen
* Sequenzen
* Dictionaries
* Klassen und Objekte
* Ausblick auf funktionales Programmieren
* Locales
* Reguläre Ausdrücke
* XML-Behandlung in Python

Leistungsnachweis
* 60% der Übungsaufgaben **müssen** erfolgreich bearbeitet werden
* Abschlussklausur
* Die erfolgreich bestandene Klausur ist Teil der Orientierungsprüfung

Formale Grundlagen der Linguistik - FF-FM, B05

V/Ü; Nr.: 09-160-02-01; SWS: 2

Mi; wöch; 16:15 - 17:45; ab 21.10.2009; INF 306 / SR 13; Hartung, M.

Kommentar

Leistungsbewertung:

FF-FM (Bachelor, neue Prüfungsordnung): 6 LP

B05 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt

Die Veranstaltung ist als Einführung in die Theorie formaler Sprachen konzipiert. Das in der Vorlesung zu erwerbende Grundwissen ist zum Verständnis der formalen Eigenschaften vieler Ansätze der Computerlinguistik zentral. Darunter fallen u.a. Grammatiktheorien in der formalen Linguistik, modelltheoretische Semantiken sowie Parsingverfahren. Insbesondere werden in der Vorlesung folgende Themen behandelt:

* Mathematische Grundlagen (Mengen, Funktionen, Relationen)
* Formale Sprachen und Grammatiken
* Reguläre Sprachen und endliche Automaten
* Kontextfreie Sprachen
* Kontextsensitive und Typ-0 Sprachen
* Turing-Maschinen
* Berechenbarkeitstheorie

Leistungsnachweis
Klausur und erfolgreiche Bearbeitung der Übungsaufgaben

Voraussetzung
Keine Voraussetzungen

Literatur
* Schöning, U.: Theoretische Informatik kurzgefasst, Spektrum, 2001
* Vossen, G. und Witt, K.-U.: Grundlagen der Theoretischen Informatik mit Anwendungen, Vieweg, 2001
* Klabunde, R.: Formale Grundlagen der Linguistik, Narr, 1998
* Partee, B. et al.: Mathematical Methods in Linguistics, Kluwer, 1990
* Hopcroft, J.E. and Ullmann, J.D.: Introduction to Automata Theory, Languages and Computation, Addison Wesley, 1979

Einführung in die statistische Sprachverarbeitung - FF-SM, A10

V/Ü; Nr.: 09-160-09-01; SWS: 2

Mi; wöch; 14:15 - 15:45; ab 21.10.2009; INF 306 / SR 13; Ponzetto, S.

Kommentar

Leistungsbewertung:

FF-SM (Bachelor, neue Prüfungsordnung): 6 LP

A10 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt

Statistische NLP-Methoden sind de-facto der Standardansatz in der aktuellen NLP-Forschung. Dieser Kurs wird eine Einführung in die theoretischen sowie in die praktischen Grundlagen der Statistischen NLP geben.

Der Schwerpunkt des Kurses wird data-driven sein, d.h. die Studierenden werden mit großen Korpora arbeiten und sie werden lernen, große Datenmengen zu handhaben. Die Anwendung von statistischen NLP-Methoden wird uns z.B. ermöglichen, Kollokationen und N-Gramme zu analysieren und diese für Textkategorisierung zu verwenden.

Wir werden uns mit einer Auswahl von bestimmten NLP-Anwendungen befassen, z.B. PoS-Tagging und Parsing, obwohl diese Methoden auf eine Vielzahl anderer NLP-Themen übertragbar sind. Als solches bietet der Kurs eine Grundlage für fortgeschrittene NLP-Themen, z.B. Maschinelle Übersetzung.

Von den Studierenden wird erwartet, dass sie ein gutes Verständnis der Theorie entwickeln und in der Lage sind, einfache NLP-Anwendungen, wie z.B. ein Hidden Markov Model oder einen Maximum Entropy basierten PoS-Tagger, zu implementieren.

Leistungsnachweis

Wöchentliche Hausaufgaben (Übungen sowie Programmieraufgaben)

Schriftliche Abschlussklausur

Zur Klausur wird nur zugelassen, wer mindestens 80% der Übungsaufgaben bearbeitet hat und mindestens 60% der maximalen Punktzahl erreicht hat.

Voraussetzung

Voraussetzung ist der erfolgreiche Abschluss der Kurse "Einführung in die Computerlinguistik" sowie "Formale Grundlagen". Programmierkenntnisse (auf dem Niveau von Programmieren I) sind für die Lösung der Übungsaufgaben von Vorteil.

Literatur

* Daniel Jurafsky and James H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Second Edition. Prentice Hall.
* Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press.
* Natural Language Toolkit: http://nltk.sourceforge.net/index.php/Book

Grundlagen Semantic Web - CS-CL, A05

V; Nr.: 09-160-10-10; SWS: 2; LP: 4

Mo; Einzel; 09:15 - 12:45; 05.10.2009 - 05.10.2009; INF 306 / SR 14; Vorlesung; Rudolph, S.

Block; 09:15 - 12:45; 06.10.2009 - 09.10.2009; INF 306 / SR 14; Vorlesung; Rudolph, S.

Block; 14:15 - 16:45; 06.10.2009 - 09.10.2009; INF 306 / SR 14; Vorlesung; Rudolph, S.

Kommentar

Leistungsbewertung:

CS-CL, BS-CL, BS-AC (Bachelor, neue Prüfungsordnung): 4 LP

A05 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt

Der Begriff Semantic Web bezeichnet allgemein eine Erweiterung des World Wide Web durch Metadaten und Anwendungen mit dem Ziel, die Bedeutung (Semantik) von Daten im Web für intelligente Systeme z.B. im E-Commerce und in Internetportalen nutzbar zu machen. Eine zentrale Rolle spielen dabei die Repräsentation und Verarbeitung von Wissen in Form von Ontologien. In dieser Vorlesung werden die Grundlagen der Wissensrepäsentation und -verarbeitung für die entsprechenden Technologien vermittelt sowie Anwendungsbeispiele vorgestellt. Dabei werden folgende Themenbereiche betrachtet:

* Grundlagen von XML (Extensible Markup Language) und XML Schema
* RDF (Resource Description Framework) und RDF Schema zur Darstellung von Metadaten und einfachen Ontologien
* Die Web Ontology Language (OWL) und ihre aktuelle Erweiterung OWL 2
* Die SPARQL-Anfragesprache für RDF, konjunktive Anfragen für OWL
* Regelsprachen für das Semantic Web
* Praktische Anwendungen

Leistungsnachweis

Leistungsnachweis durch Klausur

Literatur Literatur wird im Kurs bekannt gegeben.

Korpuslinguistik - CS-CL, A12

V; Nr.: 09-160-10-08; SWS: 2

Fr; wöch; 11:15 - 12:45; ab 23.10.2009; INF 325 / SR 24; Zielinski, A.

Kommentar Leistungsbewertung:

CS-CL, BS-CL, BS-AC (Bachelor, neue Prüfungsordnung): 4 LP (Klausur) oder 6 LP (Klausur und Referat)

A12 (Bachelor, alte Prüfungsordnung): 4 LP

Übergreifende Kompetenzen: 2 LP

Inhalt In der Korpuslinguistik werden linguistische Datensammlungen (Sprachkorpora) systematisch gesammelt und gepflegt, da sie die Basis für linguistische Forschung bilden und zur Überprüfung linguistischer Theorien dienen können. Der Begriff 'Korpus' ist definiert als "eine Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehrerer Sprachen. [...] Die Bestandteile des Korpus, die Texte oder Äußerungsfolgen, bestehen aus den Daten selbst sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind." (Lemnitzer/Zinsmeister).

In der Vorlesung geht es um den Einsatz von Korpora in unterschiedlichen Bereichen der Sprachwissenschaft. Ausgehend von den theoretischen Fragestellungen (z. B. in der computerunterstützten Lexikographie oder der Maschinellen Übersetzung) werden grundlegende korpuslinguistische Methoden vorgestellt. Dazu gehören insbesondere effiziente Technologien für die Korpussuche mit Tools wie XAIRA, Cosmas oder TigerSearch als auch Werkzeuge zur quantitativen Analyse (Kookkurrenzanalyse, Translation Memories, etc.).

Leistungsnachweis Leistungsnachweis ist eine Klausur (4 LP) oder Referat und Klausur (6 LP)

Voraussetzung Die Teilnehmerzahl für diese Veranstaltung ist begrenzt. Bei zu vielen Teilnehmern haben Studierende der Computerlinguistik Vorrang.

Literatur
* L. Lemnitzer/H. Zinsmeister, Korpuslinguistik: Eine Einführung, Narr, Tübingen 2006
* Ausgewählte Artikel aus: Anke Lüdeling & Merja Kytö (Hgg.) (erscheint 2008): Corpus Linguistics. An International Handbook. Mouton de Gruyter, Berlin.
* K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde and H. Langer (ed.): Computerlinguistik und Sprachtechnologie - Eine Einführung. Heidelberg, Spektrum-Verlag. 2001

Parsing - ACL, B09

V/Ü; Nr.: 09-160-08-01; SWS: 2

Di; wöch; 11:15 - 12:45; ab 20.10.2009; INF 327 / SR 1; Thater, S.

Kommentar Leistungsbewertung:

ACL (Bachelor, neue Prüfungsordnung): 6 LP

B09 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt Die Vorlesung stellt verschiedene Verfahren für die Syntaxanalyse (Parsing) vor. Neben klassischen Strategien (top-down, bottom-up, left-corner) und Algorithmen für kontextfreie Grammatiken werden wir Parsing-Verfahren für lexikalisierte und unifikationsbasierte Grammatikformalismen sowie stochastische Erweiterungen bestehender Ansätze betrachten.

Leistungsnachweis Klausur und erfolgreiche Bearbeitung der Übungsaufgaben

Voraussetzung Formale Grundlagen, Einführung in die Computerlinguistik

Literatur
* Naumann und Langer: Parsing. B. G. Teubner, 1994.
* Grune und Jacobs: Parsing Techniques. 2. Auflage. Springer, 2008.

* Stuart M. Shieber, Yves Schabes & Fernando C. N. Pereira (1995). Principles and implementation of deductive parsing. The Journal of Logic Programming Volume 24, Issues 1-2.

Weitere Literatur wird zu Beginn der Veranstaltung bekanntgegeben.

Maschinelle Übersetzung - CS-CL, BS-CL, BS-AC, A20

V; Nr.: 09-160-10-05; SWS: 2; LP: 4

Mo; Einzel; 09:15 - 12:45; 05.10.2009 - 05.10.2009; INF 366 / SR 12; Vorlesung; Eberle, K.

Mo; Einzel; 14:15 - 15:45; 05.10.2009 - 05.10.2009; INF 366 / SR 12; Vorlesung; Eberle, K.

Block; 09:15 - 12:45; 06.10.2009 - 09.10.2009; INF 327 / SR 4; Vorlesung; Eberle, K.

Block; 14:15 - 15:45; 06.10.2009 - 09.10.2009; INF 327 / SR 4; Vorlesung; Eberle, K.

Kommentar

Leistungsbewertung:

CS-CL, BS-CL, BS-AC (Bachelor, neue Prüfungsordnung): 4 LP

A20 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt

Nach einem kurzen Überblick über die Geschichte der Maschinellen Übersetzung werden die verschiedenen sog. regel-basierten Architekturen vorgestellt, die bis Ende der 90er Jahre die Maschinelle Übersetzung bestimmt haben (das sind vor allem die direkte Übersetzung, Transfer- und Interlingua-Verfahren). An Übersetzungsbeispielen und -schwierigkeiten werden die Vor- und Nachteile der Verfahren exemplifiziert.

Anhand der Entwicklungsumgebung des Übersetzungssystems translate wird Einblick in die Umsetzung von Spielarten der Transfer-Konzeption in einem kommerziellen System gegeben, insbesondere werden dabei Regeln aus verschiedenen System-Komponenten, wie lexikalischer Lookup, grammatische Analyse, Transfer und Generierung, exemplarisch skizziert und deren Wirkungsweise an Testbeispielen demonstriert.

Seit den 90er Jahren werden vermehrt andere, Korpus-basierte, Methoden für die Maschinelle Übersetzung diskutiert. Im zweiten Teil der Veranstaltung wird in solche Methoden, insbesondere die Grundlagen der sog. Statistik-basierten und der Beispiel-basierten Übersetzung eingeführt und am Beispiel von translate motiviert, wie Methoden kombiniert werden können.

Angesichts der zur Verfügung stehenden Zeit und der Vorkenntnisse ist das Lernziel, einen Eindruck zu vermitteln, über die Schwierigkeiten mit der eine Maschine bei der Übersetzung konfrontiert ist, über gegangene und mögliche Wege, die Aufgabe algorithmisch zu bewältigen und über die Vor- und Nachteile, die den verschiedenen Konzeptionen immanent sind.

Leistungsnachweis

Klausur

Literatur

einführende Literatur :

* Arnold, D., L. Balkan, R.L. Humphreys, S. Meijer & L. Sadler (1994): Machine Translation: An Introductory Guide, Oxford, NCC Blackwell. http:/www.essex.ac.uk/linguistics/clmt/MTbook/HTML/book.html
* Nirenburg, Sergei (ed.) (2003) Readings in Machine Translation.Cambridge: MIT Press.
* Schwanke, M. (1991): Maschinelle Übersetzung- Ein Überblick über Theorie und Praxis, Springer Verlag.
* Trujillo, A. (1999): Translation Engines: Techniques for Machine Translation, Springer Verlag.

weiterführende Literatur zu verschiedenen Methoden :

* Beaven, J. (1992): Shake and Bake Machine Translation, in COLING92.
* P. F. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, & P. Roossin, "A Statistical Approach to Machine Translation," Computational Linguistics 16(2), 1990.

* Carl, M., Way, A. (ed.) (2003): Recent Advances in Example-Based Machine Translation, Kluwer Academic Publishers, Dordrecht.
* Manning, Christopher D., Schütze, Hinrich: Chap. 13 Statistical Alignment and Machine Translation. In: Manning, Schütze: Foundations of Statistical NLP, 1999
* Michael McCord: Design of LMT, in: Computational Linguistics (15) 1989
* Sumita, E. , Iida, H., Kohyama, H.: Translating with Examples. In: A New Approach to Machine Translation. The Third International Conference on Theoretical and Methodological Issues in Machine Translation, 1990.

Information Retrieval - V01, SS-TAC, SS-CL, AS-CL

HpS; Nr.: 09-160-20-07; SWS: 2

Mo; wöch; 11:15 - 12:45; ab 19.10.2009; INF 325 / SR 24; Haenelt, K.

Kommentar — Leistungsbewertung:

AS-CL (Bachelor, neue Prüfungsordnung): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

SS-CL, SS-TAC (Master): 8 LP

Inhalt — Information Retrieval Systeme sollen Informationssuchende dabei unterstützen, aus großen elektronisch verfügbaren Informationsmengen (Texte, Datenbanken, multimediale Dokumente) passende Information herauszufinden. Im Seminar sollen die verschiedenen Ansätze und grundlegende Methoden und Algorithmen solcher Systeme erarbeitet und vermittelt werden.

Leistungsnachweis — Durchführung eines Seminarprojektes und ein Referat

Voraussetzung — Zwischenprüfung in Computerlinguistik oder vergleichbare Kenntnisse, Programmierkenntnisse (möglichst C/C++/JAVA), Programmierprüfung

Computerlinguistisches Kolloquium - Coll, V02

K; Nr.: 09-160-20-04; SWS: 2

Di; k.A.; 18:15 - 19:45; ab 13.10.2009; INF 325 / SR 24; Frank, A.

Kommentar — Leistungsbewertung:

Coll (Master): 2 LP

V02 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt — **Präsentation laufender BA-, MA- und Magisterarbeiten**

Das Computerlinguistische Kolloquium bietet BA-, MA- und Magisterstudierenden ein Forum für die Vorstellung und Diskussion ihrer Abschlussarbeiten. Die Studierenden präsentieren den aktuellen Stand ihrer Arbeit und erhalten in der Diskussion Anregungen von Seiten der Studierenden und der Dozenten.

Externe Vorträge

Darüber hinaus bietet das Computerlinguistische Kolloquium allen Studierenden durch Vorträge geladener Gäste Einblicke in aktuelle Forschungsfragen der Computerlinguistik.

Externe Vorträge finden im Rahmen des Doktorandenkolloquiums (Do, 18:15-19:45) statt.

Organisation

In den ersten beiden Sitzungen werden

* allgemeine Fragen zum Ablauf der Prüfungsphase erläutert und
* Themenvorschläge für Abschlussarbeiten vorgestellt.

Die Teilnahme an diesen Einführungssitzungen wird den **Studierenden aller Studiengangarten** , die sich vor der Prüfungsphase befinden, dringend empfohlen. Sie entlasten hierdurch die Sprechstunden.

Leistungsnachweis Vortrag (ABA, MA) und Ausarbeitung (ABA); regelmäßige Präsenz ist Voraussetzung für den Scheinerwerb.

Ein Leistungserwerb ist nur für Examenskandidat/innen im Bachelorstudiengang (ABA) und Masterstudiengang (MA) vorgesehen. Jedoch sind alle Studierenden eingeladen, ihre Abschlussarbeiten vorzustellen, den Vorträgen zuzuhören und sich an den Diskussionen zu beteiligen.

Begleitveranstaltung zum Software-Projekt - SP, V03

S; Nr.: 09-160-12-01; SWS: 2

Di; wöch; 14:15 - 15:45; ab 20.10.2009; INF 325 / SR 24; Reiter, N.

Di; wöch; 16:15 - 17:45; ab 20.10.2009; INF 325 / SR 24; Frank, A.

Kommentar Leistungsbewertung:

SP (Bachelor, neue Prüfungsordnung): 6 LP + 4 LP ÜK

V03 (Bachelor, alte Prüfungsordnung): 6 LP

Inhalt Im Softwareprojekt soll eine computerlinguistische Aufgabenstellung weitgehend eigenverantwortlich und in Teamarbeit geplant, softwaretechnisch durchgeführt, dokumentiert und abschließend präsentiert werden.

Neben der Vertiefung praktischer Programmierkenntnisse (Techniken und Werkzeuge für verteilte Programmerstellung, Testverfahren und Qualitätskontrolle, Dokumentation, etc.) sollen Teamfähigkeit und planerische Fähigkeiten geübt werden. Daneben werden grundlegende Techniken und Methoden wissenschaftlichen Arbeitens vermittelt.

Leistungsnachweis Teilnahme an allen Einführungsvorlesungen, Projekt- Spezifikationsvortrag, Projekt-Abschlussvortrag und Demo, Programmdokumentation und Archivierung

Voraussetzung Programmierprüfung, Einführung in die Benutzung computerlinguistischer Ressourcen

Voranmeldung: Per Mail an reiter@cl.uni-heidelberg.de

Literatur abhängig vom Projekt; wird zu Beginn des Semesters bekannt gegeben

Einführung in die Nutzung computerlinguistischer Ressourcen

Ü; Nr.: 09-160-00-02; SWS: 2

Block; 10:00 - 13:00; 12.10.2009 - 16.10.2009; INF 325 / PCPool; Reiter, N.

Block; 14:00 - 17:00; 12.10.2009 - 16.10.2009; INF 325 / PCPool; Reiter, N.

Kommentar
* begrenzte Teilnehmerzahl
* ggf. Vorzug für TeilnehmerInnen am Software-Projekt.

Inhalt Der Vorkurs vermittelt Grundlagen der Nutzung von Linux-basierten computerlinguistischen Tools und Korpora. Dabei geht es sowohl um allgemeine Linux-Grundlagen (wie z.B. Ein-/Ausgabeumleitung oder nützliche Tools der Linux-Kommandozeile) als auch um einzelne Parser, Tagger, Chunker und andere Hilfstools der Computerlinguistik.

Wir werden uns anschauen, wie bestimmte Tools zu benutzen sind, was man aus ihnen herausbekommt (und was nicht) und wie man solche Ausgaben automatisch weiterverarbeiten kann (und zum Beispiel an das nächste Tool weiterverfüttert).

Der Kurs beinhaltet Übungen - Wenn es nicht genug Arbeitsplätze für alle gibt, werden TeilnehmerInnen am Softwareprojekt vorgezogen.

Leistungsnachweis Ungeprüft, unbenoteter Schein

Voraussetzung Programmierprüfung

Formale Semantik - FSem, A07

V/Ü; Nr.: 09-160-07-01; SWS: 4

Di; wöch; 16:15 - 17:45; ab 20.10.2009; INF 306 / SR 19; Thater, S.

Do; wöch; 14:15 - 15:45; ab 22.10.2009; INF 306 / SR 13; Thater, S.

Kommentar

Leistungsbewertung:

FSem (Bachelor, neue Prüfungsordnung): 6 LP

A07 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt

Die Vorlesung soll einen möglichst breiten Überblick über Phänomene und Problemfelder in der Semantik natürlicher Sprachen vermitteln, die computerlinguistisch relevanten Semantikformalismen und -theorien diskutieren und Werkzeuge und Techniken für die Bedeutungsverarbeitung vorstellen.

Die Vorlesung gliedert sich grob in drei Teile: Der erste Teil vermittelt die logischen Grundlagen der modelltheoretischen (Satz-) Semantik und diskutiert Verfahren für die Semantik-Konstruktion. Wir betrachten das Phänomen der Quantifikation und stellen Verfahren für die effiziente Verarbeitung von Skopus-Ambiguitäten vor.

Der zweite Teil der Vorlesung widmet sich der formalen Behandlung von text- und diskurssemantischen Phänomenen wie Anaphorik, Koreferenz und Präsupposition am Beispiel der Diskursrepräsentationstheorie (DRT).

Im dritten Teil diskutieren wir Beschreibungsmodelle der lexikalischen Semantik (Dekomposition, Bedeutungsrelationen, Ereignisstruktur und thematische Rollen), und Modelle für die Formalisierung in Wortnetzen und Ontologien.

Voraussetzung

Foundations of Linguistic Analysis (FLA),

Formal Foundations, Logical Foundations (FF-L)

Literatur

* L.T.F. Gamut (1991). Logic, Language, and Meaning. Volume 2: Intensional Logic and Logical Grammar. The University of Chicago Press.
* Hans Kamp und Uwe Reyle (1993). From Discourse to Logic. Kluwer Academic Publishers.

Weitere Literatur wird zu Beginn der Veranstaltung bekanntgegeben.

Spracherkennung - CS-CL, BS-CL, BS-AC, A18

PS; Nr.: 09-160-10-12; SWS: 2

Fr; 14täg.; 10:15 - 13:45; ab 23.10.2009; INF 325 / PCPool; Günther, C.;Klehr, M.

Kommentar

Leistungsbewertung:

CS-CL, BS-CL, BS-AC (Bachelor, neue Prüfungsordnung): 4 LP

A18 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt

Der Kurs wird die Grundlagen der Spracherkennung behandeln. Es werden die verschiedenen Verarbeitungsschritte der automatischen Spracherkennung behandelt: von der Signalverarbeitung bis zum Sprachmodell. Dabei wird auf aktuelle Forschungen auf diesem Gebiet eingegangen. Aber auch aktuelle Implementationen und Systeme (wie der IBM WebSphere Voice Server) sollen vorgestellt werden.

Im praktischen Teil des Seminars wird auf der Grundlage von VoiceXML ein Sprachdialogsystem implementiert. Es werden die einzelnen Schritte des Entwurfs und der Implementierung behandelt (Wizard-of-Oz Test, Dialogmodell, Grammatikentwurf, Prompt-Design, Test). Es werden dabei die verschiedenen Einflussfaktoren wie Vokabulargröße oder Grammatikkomplexität auf das Erkennungsergebnis untersucht.

Leistungsnachweis

Ausarbeitung einer Programmieraufgabe (Sprachdialog-Modul)

Voraussetzung

Kenntnisse in Statistik und Signalverarbeitung sind von Vorteil, aber nicht erforderlich. Im Kurs werden Übungsaufgaben in VoiceXML gelöst, so dass Programmiererfahrungen (Java Script, XML) ebenfalls von Vorteil sind.

Literatur

* C. Günther, M. Klehr: VoiceXML 2.0, mitp 2003
* F. Jelinek: Statistical Methods for Speech Recognition, MIT Press 1997
* E. G. Schukat-Talamazzini: Automatische Spracherkennung, Vieweg 1995

* B. Eppinger, E. Herter: Sprachverarbeitung, Hanser 1993

Natural Language Generation for Virtual Environments - CS-CL, BS-CL, AC, A13

PS; Nr.: 09-160-10-22; SWS: 2

Mo; wöch; 16:15 - 17:45; ab 19.10.2009; INF 325 / SR 24; Roth, M.

Kommentar

Leistungsbewertung:

CS-CL, BS-CL, AC (Bachelor, neue Prüfungsordnung): 6 LP

A13 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt

Sprachgenerierung (auch Natural Language Generation, kurz NLG, genannt) bezeichnet ein Teilgebiet der Computerlinguistik, das sich mit der sprachlichen Realisierung aus semantischen/logischen Repräsentationen befasst. Dabei kann diese Aufgabe als komplexer Prozess verstanden werden, der aus verschiedenen Teilaufgaben wie beispielsweise Inhalts- und Diskursplanung, Wortwahl und Oberflächenrealisierung besteht.

Dieser Kurs gibt eine Einführung in die Sprachgenerierung mit dem Ziel ein eigenes Generierungssystem zu planen und zu implementieren. In den ersten Wochen des Kurses werden wir ausgewählte Publikationen zur Sprachgenerierung aufarbeiten und diskutieren, um eine Grundlage für den zweiten Kursteil zu legen. Im zweiten Teil wollen wir dann die gewonnenen Einsichten anwenden und in Gruppenarbeit ein System entwickeln, welches sprachliche Anweisungen in virtuellen Umgebungen generieren soll.

Durch die Mitarbeit im Kurs bietet sich die Möglichkeit zur Teilnahme an der GIVE-Challenge (http://www.give-challenge.org/), einem international organisierten Wettbewerb von Sprachgenerierungssystemen.

Leistungsnachweis

* Lektüre der zugrundegelegten Literatur
* Aktive und regelmäßige Teilnahme
* Gruppenprojekt und schriftliche Ausarbeitung
* Je nach Teilnehmerzahl ggf. Referat

Voraussetzung

Programmierprüfung

Data-Driven Grammar Induction - V01, AS-CL, AS-FL, SS-CL, SS-FAL

HpS; Nr.: 09-160-20-06; SWS: 2

Mi; wöch; 11:15 - 12:45; ab 21.10.2009; INF 327 / SR 1; Frank, A.

Kommentar

Leistungsbewertung:

SS-CL, SS-FAL (Master): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

Inhalt

Seit den 80/90er Jahren wurden linguistisch motivierte und formal wohldefinierte Grammatikformalismen entwickelt, insbesondere Lexical-Functional Grammar (LFG), Combinatory Categorial Grammar (CCG), Head-driven Phrase-Structure Grammar (HPSG) und Lexicalised Tree-Adjoining Grammar(LTAG). Durch die Entwicklung effizienter Parsingalgorithmen ist der Einsatz dieser Grammatikformalismen in computerlinguistischen Anwendungen realistisch geworden. Die Entwicklung umfangreicher manuell definierter Grammatiken ist zeitaufwendig und teuer; für multilinguale Sprachverarbeitung müssen jedoch umfangreiche und robuste Grammatiken in kurzer Zeit entwickelt werden.

Das Seminar führt ein in die Methodik der automatischen Induktion probabilistischer Grammatiken aus Baumbanken am Beispiel von PCFGs. Wir diskutieren insbesondere spezielle Verfahren für die automatische Induktion lexikalisierter und constraint-basierter Grammatiken (wie LFG, TAG, CCG und HPSG) aus angereicherten Baumbanken bzw. Baumbankgrammatiken. Hierbei werden wir die Charakteristiken der jeweiligen Grammatikformalismen und die entsprechenden Unterschiede der

	entsprechenden Grammatikinduktionsverfahren herausarbeiten. Abschließend widmen wir uns neueren Ansätzen für die Grammatikinduktion auf Basis paralleler Korpora.
Leistungsnachweis	Lektüre der zugrundegelegten Literatur, Referat und Hausarbeit oder Referat und Projekt
Voraussetzung	Programmierprüfung, Kenntnisse in Syntax
Literatur	* Aoife Cahill (2008): Treebank-Based Probabilistic Phrase Structure Parsing in: Language and Linguistics Compass 2/1, Blackwell, pp. 36-58. * Daniel Jurafsky and James Martin (2008): Speech and Language Processing, Kap. 13 und Kap. 14 * ESSLLI Course 2006: Josef van Genabith, Julia Hockenmaier and Yusuke Miyao: Treebank-Based Acquisition of LFG, HPSG and CCG Resources Weitere Literatur wird zu Beginn des Semesters bekanntgegeben.

Events in Discourse - V01, SS-CL, SS-FAL, AS-CL, AS-FL

HpS; Nr.: 09-160-20-14; SWS: 2

Do; wöch; 16:15 - 17:45; ab 22.10.2009; INF 325 / SR 24; Frank, A.

Kommentar	Leistungsbewertung: SS-CL, SS-FAL (Master): 8 LP V01 (Bachelor, alte Prüfungsordnung): 6 LP AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP
Inhalt	In diesem Seminar beleuchten wir vielfältige Phänomene der Semantik von Ereignissen. Im Zentrum stehen dabei die diskurssemantischen Aspekte der Verbsemantik und ihre Relevanz für automatische Diskursverarbeitung und maschinelles Textverstehen. Das Seminar behandelt zunächst die linguistischen Grundlagen sowie den Stand der Forschung zur Modellierung von Verbsemantik in computerlinguistischen Ressourcen, zu automatischen Verfahren für die Lexikonakquisition und zur automatischen Analyse von Ereignissen im Diskurs. Die Semantik von Verben und Verbklassen steht in systematischer Beziehung zu diskurssemantischen Aspekten, die konstitutiv sind für das automatische Textverstehen. Wir betrachten insbesondere: (i) implizite semantische Relationen zwischen Ereignissen bzw. Zuständen (z.B. Präsupposition, Implikation, Kausalität), (ii) temporale Relationen zwischen Ereignissen und Zuständen im Diskurs, sowie deren Lokalisierung relativ zu Zeit und Raum, (iii) die Interaktion von Verbsemantik und Diskursrelationen, wie sie vor allem in der SDRT im Vordergrund steht, bis hin zu: (iv) Phänomenen der Anaphorik. Neben der formal-linguistischen Analyse und Modellierung dieser Phänomene untersuchen wir vor allem datengetriebene Methoden für die Akquisition und die automatische Verarbeitung der Semantik von Ereignissen im Diskurs.
Leistungsnachweis	Regelmäßige Teilnahme; aktive Mitarbeit; Referat und Hausarbeit oder Projekt Vereinbarung von Referatsthemen: ab Oktober
Voraussetzung	Programmierprüfung
Literatur	Wird zu Beginn des Semesters bekanntgegeben

Unterspezifikationsformalismen für die semantische Verarbeitung - AS-CL, AS-FL, V01, SS-CL, SS-FAL

HpS; Nr.: 09-160-20-17; SWS: 2

Mo; wöch; 18:15 - 19:45; ab 12.10.2009; INF 325 / SR 24; Herweg, M.

Kommentar

Leistungsbewertung:

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

SS-CL, SS-FAL (Master): 8 LP

Inhalt

Der effiziente Umgang mit Mehrdeutigkeiten ist eine der größten Herausforderungen in der Sprachverarbeitung. Das Problem besteht darin, dass die natürliche Sprache voll von offensichtlichen oder versteckten Mehrdeutigkeiten ist und dass eine Grammatik umso mehr von diesen Mehrdeutigkeiten entdeckt, je umfassender sie die möglichen Konstruktionen einer Sprache abdeckt. Wenn nun für jede Lesart einer mehrdeutigen Konstruktion eigene Repräsentationen aufgebaut und als je eigene Analysepfade parallel oder nacheinander verfolgt werden müssen, stößt ein computerlinguistisches System schnell an seine Verarbeitungsgrenzen.

Aus diesem Grund wurde in den letzten Jahren eine Reihe von Verfahren entwickelt, die es erlauben, Mehrdeutigkeiten in einer kompakten Repräsentation darzustellen, die bezüglich der verschiedenen Lesarten unterspezifiziert ist. Erst wenn zusätzliche Information, z.B. aus dem sprachlichen Kontext oder aus der Äußerungssituation, bestimmte Lesarten ausschließt und andere favorisiert, werden die Repräsentationen sukzessive spezifischer.

Wir wollen uns in diesem Hauptseminar auf die wichtigsten Unterspezifikationsformalismen für die semantische Verarbeitung konzentrieren. Zunächst verschaffen wir uns einen Überblick über die einschlägigen linguistischen Phänomene und zentrale computerlinguistische Anwendungsbereiche für Unterspezifikation. Im Anschluss daran erarbeiten wir die wichtigsten Verfahren, die in computerlinguistischen Anwendungen eingesetzt werden.

Leistungsnachweis

Referat und schriftliche Hausarbeit (Ausarbeitung des Referats)

Voraussetzung

Logikkenntnisse (Einführung in die Logik), Grundkenntnisse in der Semantik

Die Sitzung am 12.10. wird für eine Auffrischung der Logik-Kenntnisse der TeilnehmerInnen verwendet (Schwerpunkt: modelltheoretische Semantik, Lambda-Kalkül).

Literatur

Die Literatur für die Referate und Hausarbeiten wird zum Beginn des Seminars vorgestellt. Zur Vorbereitung werden die Kapitel über Semantik in:

* Carstensen, K.U., et al. (2001): Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg/Berlin: Spektrum Akademischer Verlag (darin Kap. 3.4)
* Görz, G., et al. (2000): Handbuch der Künstlichen Intelligenz. 3. Auflage, München/Wien: Oldenbourg Verlag (darin Kap. 19) empfohlen.

Die Teilnehmer/innen sollten sich darauf einstellen, dass der größte Teil der Literatur für Referate und Hausarbeiten nur in englischer Sprache vorliegt.

Paraphrasen und Inferenz - V01, SS-CL, SS-FAL, AS-CL, AS-FL

HpS; Nr.: 09-160-20-18; SWS: 2

Mi; wöch; 14:15 - 15:45; ab 21.10.2009; INF 327 / SR 3; Thater, S.

Kommentar

Leistungsbewertung:

V01 (Bachelor, alte Prüfungsordnung): 6 LP

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

SS-CL, SS-FAL (Master): 8 LP

Inhalt

Unter Paraphrasen versteht man sprachliche Ausdrücke, die annähernd dieselbe Bedeutung haben. Dass gleiche Information durch verschiedene sprachliche Ausdrücke realisiert werden kann, ist für die Qualität und Robustheit sprachtechnologischer Anwendungen häufig ein Problem. In diesem Seminar wollen wir verschiedene

Techniken und Methoden für die automatische Identifikation und Generierung von Paraphrasen diskutieren (wobei wir von einem recht weit gefassten Paraphrasenbegriff ausgehen werden), sowie deren Anwendung und sprachverarbeitende Systeme.

Literatur wird in der ersten Sitzung bekanntgegeben

Doktoranden-Kolloquium

K; SWS: 2

Do; wöch; 18:15 - 19:45; INF 327 / SR 20; externe Vorträge; Frank, A.;Strube, M.

Inhalt Das Kolloquium bietet Doktoranden des Seminars für Computerlinguistik sowie der Abteilung NLP der EML Research gGmbH ein Forum für die Vorstellung und Diskussion ihrer laufenden Doktorarbeiten, sowie gemeinsame Lektüre und Diskussion zu ausgewählten Themenbereichen der Computerlinguistik.

Auch Bachelor- und Magisterabsolventen soll hier die Möglichkeit gegeben werden, ihre Abschlussarbeiten vorzustellen.

Im Rahmen des Kolloquiums finden externe Vorträge eingeladener Gastwissenschaftler der EML Research gGmbH und des Seminars für Computerlinguistik statt, zu denen interessierte Wissenschaftler und Studenten herzlich eingeladen sind.

Bachelor (neue Prüfungsordnung)

Einführung in die Computerlinguistik - ICL, B01

V/Ü; Nr.: 09-160-01-01; SWS: 4; LP: 6

Di; wöch; 09:15 - 10:45; ab 13.10.2009; INF 350 / OMZ R U013; Frank, A.

Do; wöch; 11:15 - 12:45; ab 22.10.2009; INF 350 / OMZ R U013; Frank, A.

Kommentar Leistungsbewertung:

ICL (Bachelor, neue Prüfungsordnung): 6 LP

B01 (Bachelor, alte Prüfungsordnung): 6 LP

Inhalt Die Vorlesung führt ein in die Grundlagen, zentralen Fragestellungen und Methoden der Computerlinguistik. In einem Gesamtüberblick werden die wesentlichen Grundlagen der Computerlinguistik eingeführt:

* Ebenen der Sprachbeschreibung (Phonologie, Morphologie, Syntax, Semantik, Pragmatik),
* formale mathematische und logische Modelle zur Beschreibung der entsprechenden linguistischen Phänomene und
* algorithmische Verfahren zur automatischen Verarbeitung auf Basis dieser Modelle.

Dabei nähern wir uns speziellen Problemen und Fragestellungen der Computerlinguistik und ihren spezifischen Lösungsstrategien. Spezielle Themen werden sein: Ambiguitätsbehandlung, Approximierung sprachlicher Regularitäten, syntaktische und semantische Verarbeitung.

Die Vorlesung gibt einen Überblick über computerlinguistische Anwendungen, diskutiert das Verhältnis zu Nachbardisziplinen, und führt durch praktische Übungen in die speziellen Fragestellungen einzelner Teilgebiete der Computerlinguistik ein.

Leistungsnachweis
* Erfolgreiche Bearbeitung der Übungsaufgaben (mind. 60%)
* Erfolgreich bestandene Klausur
* Aktive Teilnahme

Regelmäßige Präsenz ist Voraussetzung für den Scheinerwerb.

Die erfolgreich bestandene Klausur ist Teil der Orientierungsprüfung.

Literatur

* Daniel Jurafsky and James H. Martin (2000): Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall Series in Artificial Intelligence. Prentice Hall.
* Kai-Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde, Hagen Langer (Hrsg.) (2004): Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg: Spektrum, Akademischer Verlag.
* Natural Language Toolkit, NLTK: http://nltk.sourceforge.net/index.php/Book

Programmieren I - PI, B02

V/Ü; Nr.: 09-160-04-01; SWS: 4

Di; wöch; 14:15 - 15:45; ab 20.10.2009; INF 306 / SR 13; Hartung, M.

Do; wöch; 16:15 - 17:45; ab 22.10.2009; INF 328 / SR 16; Hartung, M.

Kommentar

Leistungsbewertung:

P1 (Bachelor, neue Prüfungsordnung): 6 LP

B02 (Bachelor, alte Prüfungsordnung): 6 LP

Übergreifende Kompetenzen: 3 LP

Inhalt

Ziel dieser Vorlesung ist, Studierenden einen ersten Überblick über die systematische Entwicklung von wartbaren und korrekten Programmen zu geben. Dies geschieht anhand der objektorientieren, interpretierten Sprache Python, die mit einem einfachen Objektmodell, guter Unterstützung der Modularisierung und einer reichen Bibliothek einen raschen Zugang zu modernen Programmiertechniken und zudem weitgehende Plattformunabhängigkeit bietet. Dabei wird versucht, den Stoff möglichst anhand konkreter (computerlinguistischer) Fragestellungen zu entwickeln.

Themen:

* Programmierung als Problemlösen
* Werte, Typen, Variablen
* Funktionen
* Kontrollstrukturen
* Sequenzen
* Dictionaries
* Klassen und Objekte
* Ausblick auf funktionales Programmieren
* Locales
* Reguläre Ausdrücke
* XML-Behandlung in Python

Leistungsnachweis

* 60% der Übungsaufgaben **müssen** erfolgreich bearbeitet werden
* Abschlussklausur
* Die erfolgreich bestandene Klausur ist Teil der Orientierungsprüfung

Einführung in die Sprachwissenschaft - FLA

V/Ü; Nr.: 09-160-03-01; SWS: 2; LP: 4

Mo; wöch; 16:15 - 17:45; ab 19.10.2009; INF 306 / SR 13; Witt, A.

Kommentar

Leistungsbewertung:

FLA (Bachelor, neue Prüfungsordnung): 4 LP

Inhalt

Diese Veranstaltung führt in die Grundlagen der Linguistik ein. Es werden dabei die Kernbereiche des Sprachsystems, wie Morphologie, Syntax, Semantik, Pragmatik, Phonetik und Phonologie, thematisiert.

Darüber hinaus werden Teilgebiete der Linguistik (z.B. Psycholinguistik, Korpuslinguistik, forensische Linguistik) angesprochen.

Leistungsnachweis Regelmäßige Teilnahme und aktive Mitarbeit, Lösung von Übungsaufgaben, Klausur.

Literatur
* Victoria A. Fromkin, Robert Rodman, Nina Hyams: An Introduction to Language. 7. oder 8. Auflage, Itps Thomson Learning oder Cengage Learning Services
* Hadumod Bußmann: Lexikon der Sprachwissenschaft, Kröner Verlag

Weitere Literatur wird im Seminar bekannt gegeben.

Formale Grundlagen der Linguistik - FF-FM, B05

V/Ü; Nr.: 09-160-02-01; SWS: 2

Mi; wöch; 16:15 - 17:45; ab 21.10.2009; INF 306 / SR 13; Hartung, M.

Kommentar Leistungsbewertung:

FF-FM (Bachelor, neue Prüfungsordnung): 6 LP

B05 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt

Die Veranstaltung ist als Einführung in die Theorie formaler Sprachen konzipiert. Das in der Vorlesung zu erwerbende Grundwissen ist zum Verständnis der formalen Eigenschaften vieler Ansätze der Computerlinguistik zentral. Darunter fallen u.a. Grammatiktheorien in der formalen Linguistik, modelltheoretische Semantiken sowie Parsingverfahren. Insbesondere werden in der Vorlesung folgende Themen behandelt:

* Mathematische Grundlagen (Mengen, Funktionen, Relationen)
* Formale Sprachen und Grammatiken
* Reguläre Sprachen und endliche Automaten
* Kontextfreie Sprachen
* Kontextsensitive und Typ-0 Sprachen
* Turing-Maschinen
* Berechenbarkeitstheorie

Leistungsnachweis Klausur und erfolgreiche Bearbeitung der Übungsaufgaben

Voraussetzung Keine Voraussetzungen

Literatur
* Schöning, U.: Theoretische Informatik kurzgefasst, Spektrum, 2001
* Vossen, G. und Witt, K.-U.: Grundlagen der Theoretischen Informatik mit Anwendungen, Vieweg, 2001
* Klabunde, R.: Formale Grundlagen der Linguistik, Narr, 1998
* Partee, B. et al.: Mathematical Methods in Linguistics, Kluwer, 1990
* Hopcroft, J.E. and Ullmann, J.D.: Introduction to Automata Theory, Languages and Computation, Addison Wesley, 1979

Einführung in die statistische Sprachverarbeitung - FF-SM, A10

V/Ü; Nr.: 09-160-09-01; SWS: 2

Mi; wöch; 14:15 - 15:45; ab 21.10.2009; INF 306 / SR 13; Ponzetto, S.

Kommentar Leistungsbewertung:

FF-SM (Bachelor, neue Prüfungsordnung): 6 LP

A10 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt Statistische NLP-Methoden sind de-facto der Standardansatz in der aktuellen NLP-Forschung. Dieser Kurs wird eine Einführung in die theoretischen sowie in die praktischen Grundlagen der Statistischen NLP geben.

Der Schwerpunkt des Kurses wird data-driven sein, d.h. die Studierenden werden mit großen Korpora arbeiten und sie werden lernen, große Datenmengen zu handhaben. Die Anwendung von statistischen NLP-Methoden wird uns z.B. ermöglichen, Kollokationen und N-Gramme zu analysieren und diese für Textkategorisierung zu verwenden.

Wir werden uns mit einer Auswahl von bestimmten NLP-Anwendungen befassen, z.B. PoS-Tagging und Parsing, obwohl diese Methoden auf eine Vielzahl anderer NLP-Themen übertragbar sind. Als solches bietet der Kurs eine Grundlage für fortgeschrittene NLP-Themen, z.B. Maschinelle Übersetzung.

Von den Studierenden wird erwartet, dass sie ein gutes Verständnis der Theorie entwickeln und in der Lage sind, einfache NLP-Anwendungen, wie z.B. ein Hidden Markov Model oder einen Maximum Entropy basierten PoS-Tagger, zu implementieren.

Leistungsnachweis Wöchentliche Hausaufgaben (Übungen sowie Programmieraufgaben)

Schriftliche Abschlussklausur

Zur Klausur wird nur zugelassen, wer mindestens 80% der Übungsaufgaben bearbeitet hat und mindestens 60% der maximalen Punktzahl erreicht hat.

Voraussetzung Voraussetzung ist der erfolgreiche Abschluss der Kurse "Einführung in die Computerlinguistik" sowie "Formale Grundlagen". Programmierkenntnisse (auf dem Niveau von Programmieren I) sind für die Lösung der Übungsaufgaben von Vorteil.

Literatur
* Daniel Jurafsky and James H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Second Edition. Prentice Hall.
* Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press.
* Natural Language Toolkit: http://nltk.sourceforge.net/index.php/Book

Maschinelle Übersetzung - CS-CL, BS-CL, BS-AC, A20

V; Nr.: 09-160-10-05; SWS: 2; LP: 4

Mo; Einzel; 09:15 - 12:45; 05.10.2009 - 05.10.2009; INF 366 / SR 12; Vorlesung; Eberle, K.

Mo; Einzel; 14:15 - 15:45; 05.10.2009 - 05.10.2009; INF 366 / SR 12; Vorlesung; Eberle, K.

Block; 09:15 - 12:45; 06.10.2009 - 09.10.2009; INF 327 / SR 4; Vorlesung; Eberle, K.

Block; 14:15 - 15:45; 06.10.2009 - 09.10.2009; INF 327 / SR 4; Vorlesung; Eberle, K.

Kommentar Leistungsbewertung:

CS-CL, BS-CL, BS-AC (Bachelor, neue Prüfungsordnung): 4 LP

A20 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt Nach einem kurzen Überblick über die Geschichte der Maschinellen Übersetzung werden die verschiedenen sog. regel-basierten Architekturen vorgestellt, die bis Ende der 90er Jahre die Maschinelle Übersetzung bestimmt haben (das sind vor allem die direkte Übersetzung, Transfer- und Interlingua-Verfahren). An Übersetzungsbeispielen und -schwierigkeiten werden die Vor- und Nachteile der Verfahren exemplifiziert.

Anhand der Entwicklungsumgebung des Übersetzungssystems translate wird Einblick in die Umsetzung von Spielarten der Transfer-Konzeption in einem kommerziellen System gegeben, insbesondere werden dabei Regeln aus verschiedenen System-Komponenten, wie lexikalischer Lookup, grammatische Analyse, Transfer und Generierung, exemplarisch skizziert und deren Wirkungsweise an Testbeispielen demonstriert.

Seit den 90er Jahren werden vermehrt andere, Korpus-basierte, Methoden für die Maschinelle Übersetzung diskutiert. Im zweiten Teil der Veranstaltung wird in solche Methoden, insbesondere die Grundlagen der sog. Statistik-basierten und der Beispiel-basierten Übersetzung eingeführt und am Beispiel von translate motiviert, wie Methoden kombiniert werden können.

Angesichts der zur Verfügung stehenden Zeit und der Vorkenntnisse ist das Lernziel, einen Eindruck zu vermitteln, über die Schwierigkeiten mit der eine Maschine bei der Übersetzung konfrontiert ist, über gegangene und mögliche Wege, die Aufgabe algorithmisch zu bewältigen und über die Vor- und Nachteile, die den verschiedenen Konzeptionen immanent sind.

Leistungsnachweis Klausur

Literatur **einführende Literatur** :

* Arnold, D., L. Balkan, R.L. Humphreys, S. Meijer & L. Sadler (1994): Machine Translation: An Introductory Guide, Oxford, NCC Blackwell. http:/www.essex.ac.uk/linguistics/clmt/MTbook/HTML/book.html
* Nirenburg, Sergei (ed.) (2003) Readings in Machine Translation.Cambridge: MIT Press.
* Schwanke, M. (1991): Maschinelle Übersetzung- Ein Überblick über Theorie und Praxis, Springer Verlag.
* Trujillo, A. (1999): Translation Engines: Techniques for Machine Translation, Springer Verlag.

weiterführende Literatur zu verschiedenen Methoden :

* Beaven, J. (1992): Shake and Bake Machine Translation, in COLING92.
* P. F. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, & P. Roossin, "A Statistical Approach to Machine Translation," Computational Linguistics 16(2), 1990.
* Carl, M., Way, A. (ed.) (2003): Recent Advances in Example-Based Machine Translation, Kluwer Academic Publishers, Dordrecht.
* Manning, Christopher D., Schütze, Hinrich: Chap. 13 Statistical Alignment and Machine Translation. In: Manning, Schütze: Foundations of Statistical NLP, 1999
* Michael McCord: Design of LMT, in: Computational Linguistics (15) 1989
* Sumita, E. , Iida, H., Kohyama, H.: Translating with Examples. In: A New Approach to Machine Translation. The Third International Conference on Theoretical and Methodological Issues in Machine Translation, 1990.

Korpuslinguistik - CS-CL, A12

V; Nr.: 09-160-10-08; SWS: 2

Fr; wöch; 11:15 - 12:45; ab 23.10.2009; INF 325 / SR 24; Zielinski, A.

Kommentar	Leistungsbewertung: CS-CL, BS-CL, BS-AC (Bachelor, neue Prüfungsordnung): 4 LP (Klausur) oder 6 LP (Klausur und Referat) A12 (Bachelor, alte Prüfungsordnung): 4 LP Übergreifende Kompetenzen: 2 LP
Inhalt	In der Korpuslinguistik werden linguistische Datensammlungen (Sprachkorpora) systematisch gesammelt und gepflegt, da sie die Basis für linguistische Forschung bilden und zur Überprüfung linguistischer Theorien dienen können. Der Begriff 'Korpus' ist definiert als "eine Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehrerer Sprachen. [...] Die Bestandteile des Korpus, die Texte oder Äußerungsfolgen, bestehen aus den Daten selbst sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind." (Lemnitzer/Zinsmeister). In der Vorlesung geht es um den Einsatz von Korpora in unterschiedlichen Bereichen der Sprachwissenschaft. Ausgehend von den theoretischen Fragestellungen (z. B. in der computerunterstützten Lexikographie oder der Maschinellen Übersetzung) werden grundlegende korpuslinguistische Methoden vorgestellt. Dazu gehören insbesondere effiziente Technologien für die Korpussuche mit Tools wie XAIRA, Cosmas oder TigerSearch als auch Werkzeuge zur quantitativen Analyse (Kookkurrenzanalyse, Translation Memories, etc.).
Leistungsnachweis	Leistungsnachweis ist eine Klausur (4 LP) oder Referat und Klausur (6 LP)
Voraussetzung	Die Teilnehmerzahl für diese Veranstaltung ist begrenzt. Bei zu vielen Teilnehmern haben Studierende der Computerlinguistik Vorrang.
Literatur	* L. Lemnitzer/H. Zinsmeister, Korpuslinguistik: Eine Einführung, Narr, Tübingen 2006 * Ausgewählte Artikel aus: Anke Lüdeling & Merja Kytö (Hgg.) (erscheint 2008): Corpus Linguistics. An International Handbook. Mouton de Gruyter, Berlin.

* K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde and H. Langer (ed.): Computerlinguistik und Sprachtechnologie - Eine Einführung. Heidelberg, Spektrum-Verlag. 2001

Formale Semantik - FSem, A07

V/Ü; Nr.: 09-160-07-01; SWS: 4

Di; wöch; 16:15 - 17:45; ab 20.10.2009; INF 306 / SR 19; Thater, S.

Do; wöch; 14:15 - 15:45; ab 22.10.2009; INF 306 / SR 13; Thater, S.

Kommentar

Leistungsbewertung:

FSem (Bachelor, neue Prüfungsordnung): 6 LP

A07 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt

Die Vorlesung soll einen möglichst breiten Überblick über Phänomene und Problemfelder in der Semantik natürlicher Sprachen vermitteln, die computerlinguistisch relevanten Semantikformalismen und -theorien diskutieren und Werkzeuge und Techniken für die Bedeutungsverarbeitung vorstellen.

Die Vorlesung gliedert sich grob in drei Teile: Der erste Teil vermittelt die logischen Grundlagen der modelltheoretischen (Satz-) Semantik und diskutiert Verfahren für die Semantik-Konstruktion. Wir betrachten das Phänomen der Quantifikation und stellen Verfahren für die effiziente Verarbeitung von Skopus-Ambiguitäten vor.

Der zweite Teil der Vorlesung widmet sich der formalen Behandlung von text- und diskurssemantischen Phänomenen wie Anaphorik, Koreferenz und Präsupposition am Beispiel der Diskursrepräsentationstheorie (DRT).

Im dritten Teil diskutieren wir Beschreibungsmodelle der lexikalischen Semantik (Dekomposition, Bedeutungsrelationen, Ereignisstruktur und thematische Rollen), und Modelle für die Formalisierung in Wortnetzen und Ontologien.

Voraussetzung

Foundations of Linguistic Analysis (FLA),

Formal Foundations, Logical Foundations (FF-L)

Literatur

* L.T.F. Gamut (1991). Logic, Language, and Meaning. Volume 2: Intensional Logic and Logical Grammar. The University of Chicago Press.
* Hans Kamp und Uwe Reyle (1993). From Discourse to Logic. Kluwer Academic Publishers.

Weitere Literatur wird zu Beginn der Veranstaltung bekanntgegeben.

Grundlagen Semantic Web - CS-CL, A05

V; Nr.: 09-160-10-10; SWS: 2; LP: 4

Mo; Einzel; 09:15 - 12:45; 05.10.2009 - 05.10.2009; INF 306 / SR 14; Vorlesung; Rudolph, S.

Block; 09:15 - 12:45; 06.10.2009 - 09.10.2009; INF 306 / SR 14; Vorlesung; Rudolph, S.

Block; 14:15 - 16:45; 06.10.2009 - 09.10.2009; INF 306 / SR 14; Vorlesung; Rudolph, S.

Kommentar

Leistungsbewertung:

CS-CL, BS-CL, BS-AC (Bachelor, neue Prüfungsordnung): 4 LP

A05 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt

Der Begriff Semantic Web bezeichnet allgemein eine Erweiterung des World Wide Web durch Metadaten und Anwendungen mit dem Ziel, die Bedeutung (Semantik) von Daten im Web für intelligente Systeme z.B. im E-Commerce und in Internetportalen nutzbar zu machen. Eine zentrale Rolle spielen dabei die Repräsentation und Verarbeitung von Wissen in Form von Ontologien. In dieser Vorlesung werden die Grundlagen der Wissensrepäsentation und -verarbeitung für die entsprechenden

Technologien vermittelt sowie Anwendungsbeispiele vorgestellt. Dabei werden folgende Themenbereiche betrachtet:

* Grundlagen von XML (Extensible Markup Language) und XML Schema
* RDF (Resource Description Framework) und RDF Schema zur Darstellung von Metadaten und einfachen Ontologien
* Die Web Ontology Language (OWL) und ihre aktuelle Erweiterung OWL 2
* Die SPARQL-Anfragesprache für RDF, konjunktive Anfragen für OWL
* Regelsprachen für das Semantic Web
* Praktische Anwendungen

Leistungsnachweis: Leistungsnachweis durch Klausur

Literatur: Literatur wird im Kurs bekannt gegeben.

Begleitveranstaltung zum Software-Projekt - SP, V03

S; Nr.: 09-160-12-01; SWS: 2

Di; wöch; 14:15 - 15:45; ab 20.10.2009; INF 325 / SR 24; Reiter, N.

Di; wöch; 16:15 - 17:45; ab 20.10.2009; INF 325 / SR 24; Frank, A.

Kommentar: Leistungsbewertung:

SP (Bachelor, neue Prüfungsordnung): 6 LP + 4 LP ÜK

V03 (Bachelor, alte Prüfungsordnung): 6 LP

Inhalt: Im Softwareprojekt soll eine computerlinguistische Aufgabenstellung weitgehend eigenverantwortlich und in Teamarbeit geplant, softwaretechnisch durchgeführt, dokumentiert und abschließend präsentiert werden.

Neben der Vertiefung praktischer Programmierkenntnisse (Techniken und Werkzeuge für verteilte Programmerstellung, Testverfahren und Qualitätskontrolle, Dokumentation, etc.) sollen Teamfähigkeit und planerische Fähigkeiten geübt werden. Daneben werden grundlegende Techniken und Methoden wissenschaftlichen Arbeitens vermittelt.

Leistungsnachweis: Teilnahme an allen Einführungsvorlesungen, Projekt- Spezifikationsvortrag, Projekt-Abschlussvortrag und Demo, Programmdokumentation und Archivierung

Voraussetzung: Programmierprüfung, Einführung in die Benutzung computerlinguistischer Ressourcen

Voranmeldung: Per Mail an reiter@cl.uni-heidelberg.de

Literatur: abhängig vom Projekt; wird zu Beginn des Semesters bekannt gegeben

Parsing - ACL, B09

V/Ü; Nr.: 09-160-08-01; SWS: 2

Di; wöch; 11:15 - 12:45; ab 20.10.2009; INF 327 / SR 1; Thater, S.

Kommentar: Leistungsbewertung:

ACL (Bachelor, neue Prüfungsordnung): 6 LP

B09 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt: Die Vorlesung stellt verschiedene Verfahren für die Syntaxanalyse (Parsing) vor. Neben klassischen Strategien (top-down, bottom-up, left-corner) und Algorithmen für kontextfreie Grammatiken werden wir Parsing-Verfahren für lexikalisierte und unifikationsbasierte Grammatikformalismen sowie stochastische Erweiterungen bestehender Ansätze betrachten.

Leistungsnachweis: Klausur und erfolgreiche Bearbeitung der Übungsaufgaben

Voraussetzung: Formale Grundlagen, Einführung in die Computerlinguistik

Literatur:
* Naumann und Langer: Parsing. B. G. Teubner, 1994.
* Grune und Jacobs: Parsing Techniques. 2. Auflage. Springer, 2008.

* Stuart M. Shieber, Yves Schabes & Fernando C. N. Pereira (1995). Principles and implementation of deductive parsing. The Journal of Logic Programming Volume 24, Issues 1-2.

Weitere Literatur wird zu Beginn der Veranstaltung bekanntgegeben.

Einführung in die Nutzung computerlinguistischer Ressourcen

Ü; Nr.: 09-160-00-02; SWS: 2

Block; 10:00 - 13:00; 12.10.2009 - 16.10.2009; INF 325 / PCPool; Reiter, N.

Block; 14:00 - 17:00; 12.10.2009 - 16.10.2009; INF 325 / PCPool; Reiter, N.

Kommentar
* begrenzte Teilnehmerzahl
* ggf. Vorzug für TeilnehmerInnen am Software-Projekt.

Inhalt
Der Vorkurs vermittelt Grundlagen der Nutzung von Linux-basierten computerlinguistischen Tools und Korpora. Dabei geht es sowohl um allgemeine Linux-Grundlagen (wie z.B. Ein-/Ausgabeumleitung oder nützliche Tools der Linux-Kommandozeile) als auch um einzelne Parser, Tagger, Chunker und andere Hilfstools der Computerlinguistik.

Wir werden uns anschauen, wie bestimmte Tools zu benutzen sind, was man aus ihnen herausbekommt (und was nicht) und wie man solche Ausgaben automatisch weiterverarbeiten kann (und zum Beispiel an das nächste Tool weiterverfüttert).

Der Kurs beinhaltet Übungen - Wenn es nicht genug Arbeitsplätze für alle gibt, werden TeilnehmerInnen am Softwareprojekt vorgezogen.

Leistungsnachweis
Ungeprüft, unbenoteter Schein

Voraussetzung
Programmierprüfung

Spracherkennung - CS-CL, BS-CL, BS-AC, A18

PS; Nr.: 09-160-10-12; SWS: 2

Fr; 14täg.; 10:15 - 13:45; ab 23.10.2009; INF 325 / PCPool; Günther, C.;Klehr, M.

Kommentar
Leistungsbewertung:

CS-CL, BS-CL, BS-AC (Bachelor, neue Prüfungsordnung): 4 LP

A18 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt
Der Kurs wird die Grundlagen der Spracherkennung behandeln. Es werden die verschiedenen Verarbeitungsschritte der automatischen Spracherkennung behandelt: von der Signalverarbeitung bis zum Sprachmodell. Dabei wird auf aktuelle Forschungen auf diesem Gebiet eingegangen. Aber auch aktuelle Implementationen und Systeme (wie der IBM WebSphere Voice Server) sollen vorgestellt werden.

Im praktischen Teil des Seminars wird auf der Grundlage von VoiceXML ein Sprachdialogsystem implementiert. Es werden die einzelnen Schritte des Entwurfs und der Implementierung behandelt (Wizard-of-Oz Test, Dialogmodell, Grammatikentwurf, Prompt-Design, Test). Es werden dabei die verschiedenen Einflussfaktoren wie Vokabulargröße oder Grammatikkomplexität auf das Erkennungsergebnis untersucht.

Leistungsnachweis
Ausarbeitung einer Programmieraufgabe (Sprachdialog-Modul)

Voraussetzung
Kenntnisse in Statistik und Signalverarbeitung sind von Vorteil, aber nicht erforderlich. Im Kurs werden Übungsaufgaben in VoiceXML gelöst, so dass Programmiererfahrungen (Java Script, XML) ebenfalls von Vorteil sind.

Literatur
* C. Günther, M. Klehr: VoiceXML 2.0, mitp 2003
* F. Jelinek: Statistical Methods for Speech Recognition, MIT Press 1997
* E. G. Schukat-Talamazzini: Automatische Spracherkennung, Vieweg 1995
* B. Eppinger, E. Herter: Sprachverarbeitung, Hanser 1993

Natural Language Generation for Virtual Environments - CS-CL, BS-CL, AC, A13

PS; Nr.: 09-160-10-22; SWS: 2

Mo; wöch; 16:15 - 17:45; ab 19.10.2009; INF 325 / SR 24; Roth, M.

Kommentar	Leistungsbewertung: CS-CL, BS-CL, AC (Bachelor, neue Prüfungsordnung): 6 LP A13 (Bachelor, alte Prüfungsordnung): 4 LP
Inhalt	Sprachgenerierung (auch Natural Language Generation, kurz NLG, genannt) bezeichnet ein Teilgebiet der Computerlinguistik, das sich mit der sprachlichen Realisierung aus semantischen/logischen Repräsentationen befasst. Dabei kann diese Aufgabe als komplexer Prozess verstanden werden, der aus verschiedenen Teilaufgaben wie beispielsweise Inhalts- und Diskursplanung, Wortwahl und Oberflächenrealisierung besteht. Dieser Kurs gibt eine Einführung in die Sprachgenerierung mit dem Ziel ein eigenes Generierungssystem zu planen und zu implementieren. In den ersten Wochen des Kurses werden wir ausgewählte Publikationen zur Sprachgenerierung aufarbeiten und diskutieren, um eine Grundlage für den zweiten Kursteil zu legen. Im zweiten Teil wollen wir dann die gewonnenen Einsichten anwenden und in Gruppenarbeit ein System entwickeln, welches sprachliche Anweisungen in virtuellen Umgebungen generieren soll. Durch die Mitarbeit im Kurs bietet sich die Möglichkeit zur Teilnahme an der GIVE-Challenge (http://www.give-challenge.org/), einem international organisierten Wettbewerb von Sprachgenerierungssystemen.
Leistungsnachweis	* Lektüre der zugrundegelegten Literatur * Aktive und regelmäßige Teilnahme * Gruppenprojekt und schriftliche Ausarbeitung * Je nach Teilnehmerzahl ggf. Referat
Voraussetzung	Programmierprüfung

Data-Driven Grammar Induction - V01, AS-CL, AS-FL, SS-CL, SS-FAL

HpS; Nr.: 09-160-20-06; SWS: 2

Mi; wöch; 11:15 - 12:45; ab 21.10.2009; INF 327 / SR 1; Frank, A.

Kommentar	Leistungsbewertung: SS-CL, SS-FAL (Master): 8 LP V01 (Bachelor, alte Prüfungsordnung): 6 LP AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP
Inhalt	Seit den 80/90er Jahren wurden linguistisch motivierte und formal wohldefinierte Grammatikformalismen entwickelt, insbesondere Lexical-Functional Grammar (LFG), Combinatory Categorial Grammar (CCG), Head-driven Phrase-Structure Grammar (HPSG) und Lexicalised Tree-Adjoining Grammar(LTAG). Durch die Entwicklung effizienter Parsingalgorithmen ist der Einsatz dieser Grammatikformalismen in computerlinguistischen Anwendungen realistisch geworden. Die Entwicklung umfangreicher manuell definierter Grammatiken ist zeitaufwendig und teuer; für multilinguale Sprachverarbeitung müssen jedoch umfangreiche und robuste Grammatiken in kurzer Zeit entwickelt werden. Das Seminar führt ein in die Methodik der automatischen Induktion probabilistischer Grammatiken aus Baumbanken am Beispiel von PCFGs. Wir diskutieren insbesondere spezielle Verfahren für die automatische Induktion lexikalisierter und constraint-basierter Grammatiken (wie LFG, TAG, CCG und HPSG) aus angereicherten Baumbanken bzw. Baumbankgrammatiken. Hierbei werden wir die Charakteristiken der jeweiligen Grammatikformalismen und die entsprechenden Unterschiede der entsprechenden Grammatikinduktionsverfahren herausarbeiten. Abschließend widmen wir uns neueren Ansätzen für die Grammatikinduktion auf Basis paralleler Korpora.
Leistungsnachweis	Lektüre der zugrundegelegten Literatur, Referat und Hausarbeit oder Referat und Projekt
Voraussetzung	Programmierprüfung, Kenntnisse in Syntax

Literatur

* Aoife Cahill (2008): Treebank-Based Probabilistic Phrase Structure Parsing in: Language and Linguistics Compass 2/1, Blackwell, pp. 36-58.
* Daniel Jurafsky and James Martin (2008): Speech and Language Processing, Kap. 13 und Kap. 14
* ESSLLI Course 2006: Josef van Genabith, Julia Hockenmaier and Yusuke Miyao: Treebank-Based Acquisition of LFG, HPSG and CCG Resources

Weitere Literatur wird zu Beginn des Semesters bekanntgegeben.

Information Retrieval - V01, SS-TAC, SS-CL, AS-CL

HpS; Nr.: 09-160-20-07; SWS: 2

Mo; wöch; 11:15 - 12:45; ab 19.10.2009; INF 325 / SR 24; Haenelt, K.

Kommentar

Leistungsbewertung:

AS-CL (Bachelor, neue Prüfungsordnung): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

SS-CL, SS-TAC (Master): 8 LP

Inhalt

Information Retrieval Systeme sollen Informationssuchende dabei unterstützen, aus großen elektronisch verfügbaren Informationsmengen (Texte, Datenbanken, multimediale Dokumente) passende Information herauszufinden. Im Seminar sollen die verschiedenen Ansätze und grundlegende Methoden und Algorithmen solcher Systeme erarbeitet und vermittelt werden.

Leistungsnachweis

Durchführung eines Seminarprojektes und ein Referat

Voraussetzung

Zwischenprüfung in Computerlinguistik oder vergleichbare Kenntnisse, Programmierkenntnisse (möglichst C/C++/JAVA), Programmierprüfung

Events in Discourse - V01, SS-CL, SS-FAL, AS-CL, AS-FL

HpS; Nr.: 09-160-20-14; SWS: 2

Do; wöch; 16:15 - 17:45; ab 22.10.2009; INF 325 / SR 24; Frank, A.

Kommentar

Leistungsbewertung:

SS-CL, SS-FAL (Master): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

Inhalt

In diesem Seminar beleuchten wir vielfältige Phänomene der Semantik von Ereignissen.

Im Zentrum stehen dabei die diskurssemantischen Aspekte der Verbsemantik und ihre Relevanz für automatische Diskursverarbeitung und maschinelles Textverstehen.

Das Seminar behandelt zunächst die linguistischen Grundlagen sowie den Stand der Forschung zur Modellierung von Verbsemantik in computerlinguistischen Ressourcen, zu automatischen Verfahren für die Lexikonakquisition und zur automatischen Analyse von Ereignissen im Diskurs.

Die Semantik von Verben und Verbklassen steht in systematischer Beziehung zu diskurssemantischen Aspekten, die konstitutiv sind für das automatische Textverstehen. Wir betrachten insbesondere:

(i) implizite semantische Relationen zwischen Ereignissen bzw. Zuständen (z.B. Präsupposition, Implikation, Kausalität),

(ii) temporale Relationen zwischen Ereignissen und Zuständen im Diskurs, sowie deren Lokalisierung relativ zu Zeit und Raum,

(iii) die Interaktion von Verbsemantik und Diskursrelationen, wie sie vor allem in der SDRT im Vordergrund steht, bis hin zu:

(iv) Phänomenen der Anaphorik.

Neben der formal-linguistischen Analyse und Modellierung dieser Phänomene untersuchen wir vor allem datengetriebene Methoden für die Akquisition und die automatische Verarbeitung der Semantik von Ereignissen im Diskurs.

Leistungsnachweis Regelmäßige Teilnahme; aktive Mitarbeit; Referat und Hausarbeit oder Projekt

Vereinbarung von Referatsthemen: ab Oktober
Voraussetzung Programmierprüfung
Literatur Wird zu Beginn des Semesters bekanntgegeben

Unterspezifikationsformalismen für die semantische Verarbeitung - AS-CL, AS-FL, V01, SS-CL, SS-FAL

HpS; Nr.: 09-160-20-17; SWS: 2

Mo; wöch; 18:15 - 19:45; ab 12.10.2009; INF 325 / SR 24; Herweg, M.

Kommentar Leistungsbewertung:

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

SS-CL, SS-FAL (Master): 8 LP

Inhalt Der effiziente Umgang mit Mehrdeutigkeiten ist eine der größten Herausforderungen in der Sprachverarbeitung. Das Problem besteht darin, dass die natürliche Sprache voll von offensichtlichen oder versteckten Mehrdeutigkeiten ist und dass eine Grammatik umso mehr von diesen Mehrdeutigkeiten entdeckt, je umfassender sie die möglichen Konstruktionen einer Sprache abdeckt. Wenn nun für jede Lesart einer mehrdeutigen Konstruktion eigene Repräsentationen aufgebaut und als je eigene Analysepfade parallel oder nacheinander verfolgt werden müssen, stößt ein computerlinguistisches System schnell an seine Verarbeitungsgrenzen.

Aus diesem Grund wurde in den letzten Jahren eine Reihe von Verfahren entwickelt, die es erlauben, Mehrdeutigkeiten in einer kompakten Repräsentation darzustellen, die bezüglich der verschiedenen Lesarten unterspezifiziert ist. Erst wenn zusätzliche Information, z.B. aus dem sprachlichen Kontext oder aus der Äußerungssituation, bestimmte Lesarten ausschließt und andere favorisiert, werden die Repräsentationen sukzessive spezifischer.

Wir wollen uns in diesem Hauptseminar auf die wichtigsten Unterspezifikationsformalismen für die semantische Verarbeitung konzentrieren. Zunächst verschaffen wir uns einen Überblick über die einschlägigen linguistischen Phänomene und zentrale computerlinguistische Anwendungsbereiche für Unterspezifikation. Im Anschluss daran erarbeiten wir die wichtigsten Verfahren, die in computerlinguistischen Anwendungen eingesetzt werden.

Leistungsnachweis Referat und schriftliche Hausarbeit (Ausarbeitung des Referats)
Voraussetzung Logikkenntnisse (Einführung in die Logik), Grundkenntnisse in der Semantik

Die Sitzung am 12.10. wird für eine Auffrischung der Logik-Kenntnisse der TeilnehmerInnen verwendet (Schwerpunkt: modelltheoretische Semantik, Lambda-Kalkül).

Literatur Die Literatur für die Referate und Hausarbeiten wird zum Beginn des Seminars vorgestellt. Zur Vorbereitung werden die Kapitel über Semantik in:

* Carstensen, K.U., et al. (2001): Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg/Berlin: Spektrum Akademischer Verlag (darin Kap. 3.4)
* Görz, G., et al. (2000): Handbuch der Künstlichen Intelligenz. 3. Auflage, München/Wien: Oldenbourg Verlag (darin Kap. 19) empfohlen.

Die Teilnehmer/innen sollten sich darauf einstellen, dass der größte Teil der Literatur für Referate und Hausarbeiten nur in englischer Sprache vorliegt.

Paraphrasen und Inferenz - V01, SS-CL, SS-FAL, AS-CL, AS-FL

HpS; Nr.: 09-160-20-18; SWS: 2

Mi; wöch; 14:15 - 15:45; ab 21.10.2009; INF 327 / SR 3; Thater, S.

Kommentar — Leistungsbewertung:

V01 (Bachelor, alte Prüfungsordnung): 6 LP

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

SS-CL, SS-FAL (Master): 8 LP

Inhalt — Unter Paraphrasen versteht man sprachliche Ausdrücke, die annähernd dieselbe Bedeutung haben. Dass gleiche Information durch verschiedene sprachliche Ausdrücke realisiert werden kann, ist für die Qualität und Robustheit sprachtechnologischer Anwendungen häufig ein Problem. In diesem Seminar wollen wir verschiedene Techniken und Methoden für die automatische Identifikation und Generierung von Paraphrasen diskutieren (wobei wir von einem recht weit gefassten Paraphrasenbegriff ausgehen werden), sowie deren Anwendung und sprachverarbeitende Systeme.

Literatur — wird in der ersten Sitzung bekanntgegeben

Doktoranden-Kolloquium

K; SWS: 2

Do; wöch; 18:15 - 19:45; INF 327 / SR 20; externe Vorträge; Frank, A.;Strube, M.

Inhalt — Das Kolloquium bietet Doktoranden des Seminars für Computerlinguistik sowie der Abteilung NLP der EML Research gGmbH ein Forum für die Vorstellung und Diskussion ihrer laufenden Doktorarbeiten, sowie gemeinsame Lektüre und Diskussion zu ausgewählten Themenbereichen der Computerlinguistik.

Auch Bachelor- und Magisterabsolventen soll hier die Möglichkeit gegeben werden, ihre Abschlussarbeiten vorzustellen.

Im Rahmen des Kolloquiums finden externe Vorträge eingeladener Gastwissenschaftler der EML Research gGmbH und des Seminars für Computerlinguistik statt, zu denen interessierte Wissenschaftler und Studenten herzlich eingeladen sind.

Master

Information Retrieval - V01, SS-TAC, SS-CL, AS-CL

HpS; Nr.: 09-160-20-07; SWS: 2

Mo; wöch; 11:15 - 12:45; ab 19.10.2009; INF 325 / SR 24; Haenelt, K.

Kommentar — Leistungsbewertung:

AS-CL (Bachelor, neue Prüfungsordnung): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

SS-CL, SS-TAC (Master): 8 LP

Inhalt — Information Retrieval Systeme sollen Informationssuchende dabei unterstützen, aus großen elektronisch verfügbaren Informationsmengen (Texte, Datenbanken, multimediale Dokumente) passende Information herauszufinden. Im Seminar sollen die verschiedenen Ansätze und grundlegende Methoden und Algorithmen solcher Systeme erarbeitet und vermittelt werden.

Leistungsnachweis — Durchführung eines Seminarprojektes und ein Referat

Voraussetzung — Zwischenprüfung in Computerlinguistik oder vergleichbare Kenntnisse, Programmierkenntnisse (möglichst C/C++/JAVA), Programmierprüfung

Computerlinguistisches Kolloquium - Coll, V02

K; Nr.: 09-160-20-04; SWS: 2

Di; k.A.; 18:15 - 19:45; ab 13.10.2009; INF 325 / SR 24; Frank, A.

Kommentar Leistungsbewertung:

Coll (Master): 2 LP

V02 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt **Präsentation laufender BA-, MA- und Magisterarbeiten**

Das Computerlinguistische Kolloquium bietet BA-, MA- und Magisterstudierenden ein Forum für die Vorstellung und Diskussion ihrer Abschlussarbeiten. Die Studierenden präsentieren den aktuellen Stand ihrer Arbeit und erhalten in der Diskussion Anregungen von Seiten der Studierenden und der Dozenten.

Externe Vorträge

Darüber hinaus bietet das Computerlinguistische Kolloquium allen Studierenden durch Vorträge geladener Gäste Einblicke in aktuelle Forschungsfragen der Computerlinguistik.

Externe Vorträge finden im Rahmen des Doktorandenkolloquiums (Do, 18:15-19:45) statt.

Organisation

In den ersten beiden Sitzungen werden

* allgemeine Fragen zum Ablauf der Prüfungsphase erläutert und
* Themenvorschläge für Abschlussarbeiten vorgestellt.

Die Teilnahme an diesen Einführungssitzungen wird den **Studierenden aller Studiengangarten** , die sich vor der Prüfungsphase befinden, dringend empfohlen. Sie entlasten hierdurch die Sprechstunden.

Leistungsnachweis Vortrag (ABA, MA) und Ausarbeitung (ABA); regelmäßige Präsenz ist Voraussetzung für den Scheinerwerb.

Ein Leistungserwerb ist nur für Examenskandidat/innen im Bachelorstudiengang (ABA) und Masterstudiengang (MA) vorgesehen. Jedoch sind alle Studierenden eingeladen, ihre Abschlussarbeiten vorzustellen, den Vorträgen zuzuhören und sich an den Diskussionen zu beteiligen.

Machine Learning - SS-CL, SS-TAC

HpS/Ü; Nr.: 09-160-20-03; SWS: 3

Do; wöch; 11:15 - 12:00; ab 22.10.2009; INF 325 / SR 24; Fendrich, S.

Do; wöch; 14:15 - 15:45; ab 22.10.2009; INF 325 / SR 24; Fendrich, S.

Kommentar Leistungsbewertung:

SS-CL, SS-TAC (Master): 8 LP

Inhalt Die Veranstaltung hat in der ersten Semesterhälfte die Form einer Vorlesung; in der zweiten Hälfte erfolgen Referate der Teilnehmer. Gegenstand der Veranstaltung sind grundlegende Methoden des Maschinellen Lernens. Behandelt werden u.a. Entscheidungsbäume, Clustering-Verfahren, Bayessches Lernen, Kernel-basierte Methoden und Support-Vector-Maschinen. Darüber hinaus wird es im Rahmen einer Übung die Gelegenheit geben, die Data-Mining-Software WEKA kennen zu lernen.

Leistungsnachweis - Bearbeitung der Übungsaufgaben

- Referat (40%)

- mündliche Prüfung (60%)

Voraussetzung - Programmierprüfung

- Formale Grundlagen oder Mathematischer Vorkurs

- Grundkenntnisse in Statistik

Literatur
* Mitchell: Machine Learning. McGraw-Hill, 1997.
* Bishop: Pattern Recognition and Machine Learning. Springer, 2006.
* Witten/Frank: Data Mining. Morgan Kaufman, 2005.

weitere Literatur wird im Kurs bekannt gegeben

Data-Driven Grammar Induction - V01, AS-CL, AS-FL, SS-CL, SS-FAL

HpS; Nr.: 09-160-20-06; SWS: 2

Mi; wöch; 11:15 - 12:45; ab 21.10.2009; INF 327 / SR 1; Frank, A.

Kommentar

Leistungsbewertung:

SS-CL, SS-FAL (Master): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

Inhalt

Seit den 80/90er Jahren wurden linguistisch motivierte und formal wohldefinierte Grammatikformalismen entwickelt, insbesondere Lexical-Functional Grammar (LFG), Combinatory Categorial Grammar (CCG), Head-driven Phrase-Structure Grammar (HPSG) und Lexicalised Tree-Adjoining Grammar(LTAG). Durch die Entwicklung effizienter Parsingalgorithmen ist der Einsatz dieser Grammatikformalismen in computerlinguistischen Anwendungen realistisch geworden. Die Entwicklung umfangreicher manuell definierter Grammatiken ist zeitaufwendig und teuer; für multilinguale Sprachverarbeitung müssen jedoch umfangreiche und robuste Grammatiken in kurzer Zeit entwickelt werden.

Das Seminar führt ein in die Methodik der automatischen Induktion probabilistischer Grammatiken aus Baumbanken am Beispiel von PCFGs. Wir diskutieren insbesondere spezielle Verfahren für die automatische Induktion lexikalisierter und constraint-basierter Grammatiken (wie LFG, TAG, CCG und HPSG) aus angereicherten Baumbanken bzw. Baumbankgrammatiken. Hierbei werden wir die Charakteristiken der jeweiligen Grammatikformalismen und die entsprechenden Unterschiede der entsprechenden Grammatikinduktionsverfahren herausarbeiten. Abschließend widmen wir uns neueren Ansätzen für die Grammatikinduktion auf Basis paralleler Korpora.

Leistungsnachweis

Lektüre der zugrundegelegten Literatur, Referat und Hausarbeit oder Referat und Projekt

Voraussetzung

Programmierprüfung, Kenntnisse in Syntax

Literatur

* Aoife Cahill (2008): Treebank-Based Probabilistic Phrase Structure Parsing in: Language and Linguistics Compass 2/1, Blackwell, pp. 36-58.
* Daniel Jurafsky and James Martin (2008): Speech and Language Processing, Kap. 13 und Kap. 14
* ESSLLI Course 2006: Josef van Genabith, Julia Hockenmaier and Yusuke Miyao: Treebank-Based Acquisition of LFG, HPSG and CCG Resources

Weitere Literatur wird zu Beginn des Semesters bekanntgegeben.

Events in Discourse - V01, SS-CL, SS-FAL, AS-CL, AS-FL

HpS; Nr.: 09-160-20-14; SWS: 2

Do; wöch; 16:15 - 17:45; ab 22.10.2009; INF 325 / SR 24; Frank, A.

Kommentar

Leistungsbewertung:

SS-CL, SS-FAL (Master): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

Inhalt

In diesem Seminar beleuchten wir vielfältige Phänomene der Semantik von Ereignissen.

Im Zentrum stehen dabei die diskurssemantischen Aspekte der Verbsemantik und ihre Relevanz für automatische Diskursverarbeitung und maschinelles Textverstehen.

Das Seminar behandelt zunächst die linguistischen Grundlagen sowie den Stand der Forschung zur Modellierung von Verbsemantik in computerlinguistischen Ressourcen,

zu automatischen Verfahren für die Lexikonakquisition und zur automatischen Analyse von Ereignissen im Diskurs.

Die Semantik von Verben und Verbklassen steht in systematischer Beziehung zu diskurssemantischen Aspekten, die konstitutiv sind für das automatische Textverstehen. Wir betrachten insbesondere:

(i) implizite semantische Relationen zwischen Ereignissen bzw. Zuständen (z.B. Präsupposition, Implikation, Kausalität),

(ii) temporale Relationen zwischen Ereignissen und Zuständen im Diskurs, sowie deren Lokalisierung relativ zu Zeit und Raum,

(iii) die Interaktion von Verbsemantik und Diskursrelationen, wie sie vor allem in der SDRT im Vordergrund steht, bis hin zu:

(iv) Phänomenen der Anaphorik.

Neben der formal-linguistischen Analyse und Modellierung dieser Phänomene untersuchen wir vor allem datengetriebene Methoden für die Akquisition und die automatische Verarbeitung der Semantik von Ereignissen im Diskurs.

Leistungsnachweis Regelmäßige Teilnahme; aktive Mitarbeit; Referat und Hausarbeit oder Projekt

Vereinbarung von Referatsthemen: ab Oktober

Voraussetzung Programmierprüfung

Literatur Wird zu Beginn des Semesters bekanntgegeben

Unterspezifikationsformalismen für die semantische Verarbeitung - AS-CL, AS-FL, V01, SS-CL, SS-FAL

HpS; Nr.: 09-160-20-17; SWS: 2

Mo; wöch; 18:15 - 19:45; ab 12.10.2009; INF 325 / SR 24; Herweg, M.

Kommentar Leistungsbewertung:

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

SS-CL, SS-FAL (Master): 8 LP

Inhalt Der effiziente Umgang mit Mehrdeutigkeiten ist eine der größten Herausforderungen in der Sprachverarbeitung. Das Problem besteht darin, dass die natürliche Sprache voll von offensichtlichen oder versteckten Mehrdeutigkeiten ist und dass eine Grammatik umso mehr von diesen Mehrdeutigkeiten entdeckt, je umfassender sie die möglichen Konstruktionen einer Sprache abdeckt. Wenn nun für jede Lesart einer mehrdeutigen Konstruktion eigene Repräsentationen aufgebaut und als je eigene Analysepfade parallel oder nacheinander verfolgt werden müssen, stößt ein computerlinguistisches System schnell an seine Verarbeitungsgrenzen.

Aus diesem Grund wurde in den letzten Jahren eine Reihe von Verfahren entwickelt, die es erlauben, Mehrdeutigkeiten in einer kompakten Repräsentation darzustellen, die bezüglich der verschiedenen Lesarten unterspezifiziert ist. Erst wenn zusätzliche Information, z.B. aus dem sprachlichen Kontext oder aus der Äußerungssituation, bestimmte Lesarten ausschließt und andere favorisiert, werden die Repräsentationen sukzessive spezifischer.

Wir wollen uns in diesem Hauptseminar auf die wichtigsten Unterspezifikationsformalismen für die semantische Verarbeitung konzentrieren. Zunächst verschaffen wir uns einen Überblick über die einschlägigen linguistischen Phänomene und zentrale computerlinguistische Anwendungsbereiche für Unterspezifikation. Im Anschluss daran erarbeiten wir die wichtigsten Verfahren, die in computerlinguistischen Anwendungen eingesetzt werden.

Leistungsnachweis Referat und schriftliche Hausarbeit (Ausarbeitung des Referats)

Voraussetzung Logikkenntnisse (Einführung in die Logik), Grundkenntnisse in der Semantik

Die Sitzung am 12.10. wird für eine Auffrischung der Logik-Kenntnisse der TeilnehmerInnen verwendet (Schwerpunkt: modelltheoretische Semantik, Lambda-Kalkül).

Literatur

Die Literatur für die Referate und Hausarbeiten wird zum Beginn des Seminars vorgestellt. Zur Vorbereitung werden die Kapitel über Semantik in:

* Carstensen, K.U., et al. (2001): Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg/Berlin: Spektrum Akademischer Verlag (darin Kap. 3.4)
* Görz, G., et al. (2000): Handbuch der Künstlichen Intelligenz. 3. Auflage, München/Wien: Oldenbourg Verlag (darin Kap. 19) empfohlen.

Die Teilnehmer/innen sollten sich darauf einstellen, dass der größte Teil der Literatur für Referate und Hausarbeiten nur in englischer Sprache vorliegt.

Paraphrasen und Inferenz - V01, SS-CL, SS-FAL, AS-CL, AS-FL

HpS; Nr.: 09-160-20-18; SWS: 2

Mi; wöch; 14:15 - 15:45; ab 21.10.2009; INF 327 / SR 3; Thater, S.

Kommentar

Leistungsbewertung:

V01 (Bachelor, alte Prüfungsordnung): 6 LP

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

SS-CL, SS-FAL (Master): 8 LP

Inhalt

Unter Paraphrasen versteht man sprachliche Ausdrücke, die annähernd dieselbe Bedeutung haben. Dass gleiche Information durch verschiedene sprachliche Ausdrücke realisiert werden kann, ist für die Qualität und Robustheit sprachtechnologischer Anwendungen häufig ein Problem. In diesem Seminar wollen wir verschiedene Techniken und Methoden für die automatische Identifikation und Generierung von Paraphrasen diskutieren (wobei wir von einem recht weit gefassten Paraphrasenbegriff ausgehen werden), sowie deren Anwendung und sprachverarbeitende Systeme.

Literatur

wird in der ersten Sitzung bekanntgegeben

Word Sense Disambiguation - V01, SS-CL

HpS; Nr.: 09-160-20-19; SWS: 2; LP: 8

Di; wöch; 09:15 - 10:45; ab 20.10.2009; INF 325 / SR 24; Ponzetto, S.

Inhalt

Word Sense Disambiguation (WSD) is the problem of identifying the intended meaning (or sense) of a word, based on the context in which it occurs. Correctly identifying the senses of words in context is a central problem for Natural Language Processing (NLP), and robust performance on this task is accordingly expected to provide crucial lexical semantic information for many NLP applications such as machine translation, information retrieval, etc.

This seminar will provide a gentle introduction to state-of-the-art approaches in WSD. These include:

* knowledge-based methods that either (a) make use of dictionaries and thesauri and/or (b) manually crafted graph-like resources such as e.g. WordNet or GermaNet;
* supervised machine learning methods that learn classifiers from sense annotated data;
* minimally supervised methods (aka bootstrapping) that, starting with a small amounts of labeled data (seeds), iteratively harvest new sense annotations to improve the sense disambiguation accuracy.

Students will present current work from the literature in short, seminar-format presentations (Referate). In addition, every 4-6 weeks they will be expected to form small groups of 3-4 people and work on a project, e.g. implement and/or extend an existing state-of-the-art WSD approach. Each of the groups is expected to submit a

short report (2-4 pages), as well as to give a short project-overview presentation at the end of each round. Students are expected to *actively* participate in the class discussions during their fellow students' presentations, as well as in the seminar's projects. This means that you'll have to read the papers **before** the class period in which they will be presented and discussed, as well as **clearly** present to the audience what your specific work was as part of the seminar's projects.

Determination of final grade:

33%: presentation

33%: participation in the seminar's projects

33%: participation in the class discussions

Leistungsnachweis: Aktive Teilnahme und regelmäßige Abgabe von Projektarbeit in kleinen Gruppen. Vortrag/Präsentation.

Voraussetzung: Voraussetzungen sind die bestandene Zwischenprüfung (Magister) und Programmierprüfung. Vorkenntnisse in statistischer NLP oder Maschinellem Lernen sind von Vorteil.

Literatur:

Vollständige Lektüre von Navigli (2009) (siehe unten)

* R. Navigli. Word Sense Disambiguation: a Survey, ACM Computing Surveys, 41(2), ACM Press, 2009, pp. 1-69 (WICHTIG: Lektüre zu Semesterbeginn vorausgesetzt!); Link: http://www.dsi.uniroma1.it/~navigli/pubs/ACM_Survey_2009_Navigli.pdf
* Eneko Agirre & Philip Edmonds (eds.) Word Sense Disambiguation Algorithms and Applications, Springer, 2006 (wird als Referenz benutzt) Link: http://www.wsdbook.org/

Doktoranden-Kolloquium

K; SWS: 2

Do; wöch; 18:15 - 19:45; INF 327 / SR 20; externe Vorträge; Frank, A.;Strube, M.

Inhalt: Das Kolloquium bietet Doktoranden des Seminars für Computerlinguistik sowie der Abteilung NLP der EML Research gGmbH ein Forum für die Vorstellung und Diskussion ihrer laufenden Doktorarbeiten, sowie gemeinsame Lektüre und Diskussion zu ausgewählten Themenbereichen der Computerlinguistik.

Auch Bachelor- und Magisterabsolventen soll hier die Möglichkeit gegeben werden, ihre Abschlussarbeiten vorzustellen.

Im Rahmen des Kolloquiums finden externe Vorträge eingeladener Gastwissenschaftler der EML Research gGmbH und des Seminars für Computerlinguistik statt, zu denen interessierte Wissenschaftler und Studenten herzlich eingeladen sind.

Magister

Einführung in die Nutzung computerlinguistischer Ressourcen

Ü; Nr.: 09-160-00-02; SWS: 2

Block; 10:00 - 13:00; 12.10.2009 - 16.10.2009; INF 325 / PCPool; Reiter, N.

Block; 14:00 - 17:00; 12.10.2009 - 16.10.2009; INF 325 / PCPool; Reiter, N.

Kommentar:

* begrenzte Teilnehmerzahl
* ggf. Vorzug für TeilnehmerInnen am Software-Projekt.

Inhalt: Der Vorkurs vermittelt Grundlagen der Nutzung von Linux-basierten computerlinguistischen Tools und Korpora. Dabei geht es sowohl um allgemeine Linux-Grundlagen (wie z.B. Ein-/Ausgabeumleitung oder nützliche Tools der Linux-Kommandozeile) als auch um einzelne Parser, Tagger, Chunker und andere Hilfstools der Computerlinguistik.

Wir werden uns anschauen, wie bestimmte Tools zu benutzen sind, was man aus ihnen herausbekommt (und was nicht) und wie man solche Ausgaben automatisch weiterverarbeiten kann (und zum Beispiel an das nächste Tool weiterverfüttert).

Der Kurs beinhaltet Übungen - Wenn es nicht genug Arbeitsplätze für alle gibt, werden TeilnehmerInnen am Softwareprojekt vorgezogen.

Leistungsnachweis Ungeprüft, unbenoteter Schein
Voraussetzung Programmierprüfung

Computerlinguistisches Kolloquium - Coll, V02

K; Nr.: 09-160-20-04; SWS: 2

Di; k.A.; 18:15 - 19:45; ab 13.10.2009; INF 325 / SR 24; Frank, A.

Kommentar Leistungsbewertung:

Coll (Master): 2 LP

V02 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt **Präsentation laufender BA-, MA- und Magisterarbeiten**

Das Computerlinguistische Kolloquium bietet BA-, MA- und Magisterstudierenden ein Forum für die Vorstellung und Diskussion ihrer Abschlussarbeiten. Die Studierenden präsentieren den aktuellen Stand ihrer Arbeit und erhalten in der Diskussion Anregungen von Seiten der Studierenden und der Dozenten.

Externe Vorträge

Darüber hinaus bietet das Computerlinguistische Kolloquium allen Studierenden durch Vorträge geladener Gäste Einblicke in aktuelle Forschungsfragen der Computerlinguistik.

Externe Vorträge finden im Rahmen des Doktorandenkolloquiums (Do, 18:15-19:45) statt.

Organisation

In den ersten beiden Sitzungen werden

* allgemeine Fragen zum Ablauf der Prüfungsphase erläutert und
* Themenvorschläge für Abschlussarbeiten vorgestellt.

Die Teilnahme an diesen Einführungssitzungen wird den **Studierenden aller Studiengangarten** , die sich vor der Prüfungsphase befinden, dringend empfohlen. Sie entlasten hierdurch die Sprechstunden.

Leistungsnachweis Vortrag (ABA, MA) und Ausarbeitung (ABA); regelmäßige Präsenz ist Voraussetzung für den Scheinerwerb.

Ein Leistungserwerb ist nur für Examenskandidat/innen im Bachelorstudiengang (ABA) und Masterstudiengang (MA) vorgesehen. Jedoch sind alle Studierenden eingeladen, ihre Abschlussarbeiten vorzustellen, den Vorträgen zuzuhören und sich an den Diskussionen zu beteiligen.

Doktoranden-Kolloquium

K; SWS: 2

Do; wöch; 18:15 - 19:45; INF 327 / SR 20; externe Vorträge; Frank, A.;Strube, M.

Inhalt Das Kolloquium bietet Doktoranden des Seminars für Computerlinguistik sowie der Abteilung NLP der EML Research gGmbH ein Forum für die Vorstellung und Diskussion ihrer laufenden Doktorarbeiten, sowie gemeinsame Lektüre und Diskussion zu ausgewählten Themenbereichen der Computerlinguistik.

Auch Bachelor- und Magisterabsolventen soll hier die Möglichkeit gegeben werden, ihre Abschlussarbeiten vorzustellen.

Im Rahmen des Kolloquiums finden externe Vorträge eingeladener Gastwissenschaftler der EML Research gGmbH und des Seminars für Computerlinguistik statt, zu denen interessierte Wissenschaftler und Studenten herzlich eingeladen sind.

Informatik und Programmierpraxis

Programmieren I - PI, B02

V/Ü; Nr.: 09-160-04-01; SWS: 4

Di; wöch; 14:15 - 15:45; ab 20.10.2009; INF 306 / SR 13; Hartung, M.

Do; wöch; 16:15 - 17:45; ab 22.10.2009; INF 328 / SR 16; Hartung, M.

Kommentar

Leistungsbewertung:

P1 (Bachelor, neue Prüfungsordnung): 6 LP

B02 (Bachelor, alte Prüfungsordnung): 6 LP

Übergreifende Kompetenzen: 3 LP

Inhalt

Ziel dieser Vorlesung ist, Studierenden einen ersten Überblick über die systematische Entwicklung von wartbaren und korrekten Programmen zu geben. Dies geschieht anhand der objektorientieren, interpretierten Sprache Python, die mit einem einfachen Objektmodell, guter Unterstützung der Modularisierung und einer reichen Bibliothek einen raschen Zugang zu modernen Programmiertechniken und zudem weitgehende Plattformunabhängigkeit bietet. Dabei wird versucht, den Stoff möglichst anhand konkreter (computerlinguistischer) Fragestellungen zu entwickeln.

Themen:

* Programmierung als Problemlösen
* Werte, Typen, Variablen
* Funktionen
* Kontrollstrukturen
* Sequenzen
* Dictionaries
* Klassen und Objekte
* Ausblick auf funktionales Programmieren
* Locales
* Reguläre Ausdrücke
* XML-Behandlung in Python

Leistungsnachweis

* 60% der Übungsaufgaben **müssen** erfolgreich bearbeitet werden
* Abschlussklausur
* Die erfolgreich bestandene Klausur ist Teil der Orientierungsprüfung

Begleitveranstaltung zum Software-Projekt - SP, V03

S; Nr.: 09-160-12-01; SWS: 2

Di; wöch; 14:15 - 15:45; ab 20.10.2009; INF 325 / SR 24; Reiter, N.

Di; wöch; 16:15 - 17:45; ab 20.10.2009; INF 325 / SR 24; Frank, A.

Kommentar

Leistungsbewertung:

SP (Bachelor, neue Prüfungsordnung): 6 LP + 4 LP ÜK

V03 (Bachelor, alte Prüfungsordnung): 6 LP

Inhalt

Im Softwareprojekt soll eine computerlinguistische Aufgabenstellung weitgehend eigenverantwortlich und in Teamarbeit geplant, softwaretechnisch durchgeführt, dokumentiert und abschließend präsentiert werden.

Neben der Vertiefung praktischer Programmierkenntnisse (Techniken und Werkzeuge für verteilte Programmerstellung, Testverfahren und Qualitätskontrolle, Dokumentation, etc.) sollen Teamfähigkeit und planerische Fähigkeiten geübt werden. Daneben werden grundlegende Techniken und Methoden wissenschaftlichen Arbeitens vermittelt.

Leistungsnachweis: Teilnahme an allen Einführungsvorlesungen, Projekt- Spezifikationsvortrag, Projekt-Abschlussvortrag und Demo, Programmdokumentation und Archivierung

Voraussetzung: Programmierprüfung, Einführung in die Benutzung computerlinguistischer Ressourcen

Voranmeldung: Per Mail an reiter@cl.uni-heidelberg.de

Literatur: abhängig vom Projekt; wird zu Beginn des Semesters bekannt gegeben

Theoretische und empirische Grundlagen der Linguistik

Formale Grundlagen der Linguistik - FF-FM, B05

V/Ü; Nr.: 09-160-02-01; SWS: 2

Mi; wöch; 16:15 - 17:45; ab 21.10.2009; INF 306 / SR 13; Hartung, M.

Kommentar: Leistungsbewertung:

FF-FM (Bachelor, neue Prüfungsordnung): 6 LP

B05 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt

Die Veranstaltung ist als Einführung in die Theorie formaler Sprachen konzipiert. Das in der Vorlesung zu erwerbende Grundwissen ist zum Verständnis der formalen Eigenschaften vieler Ansätze der Computerlinguistik zentral. Darunter fallen u.a. Grammatiktheorien in der formalen Linguistik, modelltheoretische Semantiken sowie Parsingverfahren. Insbesondere werden in der Vorlesung folgende Themen behandelt:

* Mathematische Grundlagen (Mengen, Funktionen, Relationen)
* Formale Sprachen und Grammatiken
* Reguläre Sprachen und endliche Automaten
* Kontextfreie Sprachen
* Kontextsensitive und Typ-0 Sprachen
* Turing-Maschinen
* Berechenbarkeitstheorie

Leistungsnachweis: Klausur und erfolgreiche Bearbeitung der Übungsaufgaben

Voraussetzung: Keine Voraussetzungen

Literatur:

* Schöning, U.: Theoretische Informatik kurzgefasst, Spektrum, 2001
* Vossen, G. und Witt, K.-U.: Grundlagen der Theoretischen Informatik mit Anwendungen, Vieweg, 2001
* Klabunde, R.: Formale Grundlagen der Linguistik, Narr, 1998
* Partee, B. et al.: Mathematical Methods in Linguistics, Kluwer, 1990
* Hopcroft, J.E. and Ullmann, J.D.: Introduction to Automata Theory, Languages and Computation, Addison Wesley, 1979

Einführung in die Sprachwissenschaft - FLA

V/Ü; Nr.: 09-160-03-01; SWS: 2; LP: 4

Mo; wöch; 16:15 - 17:45; ab 19.10.2009; INF 306 / SR 13; Witt, A.

Kommentar: Leistungsbewertung:

FLA (Bachelor, neue Prüfungsordnung): 4 LP

Inhalt
Diese Veranstaltung führt in die Grundlagen der Linguistik ein. Es werden dabei die Kernbereiche des Sprachsystems, wie Morphologie, Syntax, Semantik, Pragmatik, Phonetik und Phonologie, thematisiert.

Darüber hinaus werden Teilgebiete der Linguistik (z.B. Psycholinguistik, Korpuslinguistik, forensische Linguistik) angesprochen.

Leistungsnachweis
Regelmäßige Teilnahme und aktive Mitarbeit, Lösung von Übungsaufgaben, Klausur.

Literatur
* Victoria A. Fromkin, Robert Rodman, Nina Hyams: An Introduction to Language. 7. oder 8. Auflage, Itps Thomson Learning oder Cengage Learning Services
* Hadumod Bußmann: Lexikon der Sprachwissenschaft, Kröner Verlag

Weitere Literatur wird im Seminar bekannt gegeben.

Formale Semantik - FSem, A07

V/Ü; Nr.: 09-160-07-01; SWS: 4

Di; wöch; 16:15 - 17:45; ab 20.10.2009; INF 306 / SR 19; Thater, S.

Do; wöch; 14:15 - 15:45; ab 22.10.2009; INF 306 / SR 13; Thater, S.

Kommentar
Leistungsbewertung:

FSem (Bachelor, neue Prüfungsordnung): 6 LP

A07 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt
Die Vorlesung soll einen möglichst breiten Überblick über Phänomene und Problemfelder in der Semantik natürlicher Sprachen vermitteln, die computerlinguistisch relevanten Semantikformalismen und -theorien diskutieren und Werkzeuge und Techniken für die Bedeutungsverarbeitung vorstellen.

Die Vorlesung gliedert sich grob in drei Teile: Der erste Teil vermittelt die logischen Grundlagen der modelltheoretischen (Satz-) Semantik und diskutiert Verfahren für die Semantik-Konstruktion. Wir betrachten das Phänomen der Quantifikation und stellen Verfahren für die effiziente Verarbeitung von Skopus-Ambiguitäten vor.

Der zweite Teil der Vorlesung widmet sich der formalen Behandlung von text- und diskurssemantischen Phänomenen wie Anaphorik, Koreferenz und Präsupposition am Beispiel der Diskursrepräsentationstheorie (DRT).

Im dritten Teil diskutieren wir Beschreibungsmodelle der lexikalischen Semantik (Dekomposition, Bedeutungsrelationen, Ereignisstruktur und thematische Rollen), und Modelle für die Formalisierung in Wortnetzen und Ontologien.

Voraussetzung
Foundations of Linguistic Analysis (FLA),

Formal Foundations, Logical Foundations (FF-L)

Literatur
* L.T.F. Gamut (1991). Logic, Language, and Meaning. Volume 2: Intensional Logic and Logical Grammar. The University of Chicago Press.
* Hans Kamp und Uwe Reyle (1993). From Discourse to Logic. Kluwer Academic Publishers.

Weitere Literatur wird zu Beginn der Veranstaltung bekanntgegeben.

Korpuslinguistik - CS-CL, A12

V; Nr.: 09-160-10-08; SWS: 2

Fr; wöch; 11:15 - 12:45; ab 23.10.2009; INF 325 / SR 24; Zielinski, A.

Kommentar
Leistungsbewertung:

CS-CL, BS-CL, BS-AC (Bachelor, neue Prüfungsordnung): 4 LP (Klausur) oder 6 LP (Klausur und Referat)

A12 (Bachelor, alte Prüfungsordnung): 4 LP

Übergreifende Kompetenzen: 2 LP

Inhalt

In der Korpuslinguistik werden linguistische Datensammlungen (Sprachkorpora) systematisch gesammelt und gepflegt, da sie die Basis für linguistische Forschung bilden und zur Überprüfung linguistischer Theorien dienen können. Der Begriff 'Korpus' ist definiert als "eine Sammlung schriftlicher oder gesprochener Äußerungen in einer oder mehrerer Sprachen. [...] Die Bestandteile des Korpus, die Texte oder Äußerungsfolgen, bestehen aus den Daten selbst sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind." (Lemnitzer/Zinsmeister).

In der Vorlesung geht es um den Einsatz von Korpora in unterschiedlichen Bereichen der Sprachwissenschaft. Ausgehend von den theoretischen Fragestellungen (z. B. in der computerunterstützten Lexikographie oder der Maschinellen Übersetzung) werden grundlegende korpuslinguistische Methoden vorgestellt. Dazu gehören insbesondere effiziente Technologien für die Korpussuche mit Tools wie XAIRA, Cosmas oder TigerSearch als auch Werkzeuge zur quantitativen Analyse (Kookkurrenzanalyse, Translation Memories, etc.).

Leistungsnachweis

Leistungsnachweis ist eine Klausur (4 LP) oder Referat und Klausur (6 LP)

Voraussetzung

Die Teilnehmerzahl für diese Veranstaltung ist begrenzt. Bei zu vielen Teilnehmern haben Studierende der Computerlinguistik Vorrang.

Literatur

* L. Lemnitzer/H. Zinsmeister, Korpuslinguistik: Eine Einführung, Narr, Tübingen 2006
* Ausgewählte Artikel aus: Anke Lüdeling & Merja Kytö (Hgg.) (erscheint 2008): Corpus Linguistics. An International Handbook. Mouton de Gruyter, Berlin.
* K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde and H. Langer (ed.): Computerlinguistik und Sprachtechnologie - Eine Einführung. Heidelberg, Spektrum-Verlag. 2001

Events in Discourse - V01, SS-CL, SS-FAL, AS-CL, AS-FL

HpS; Nr.: 09-160-20-14; SWS: 2

Do; wöch; 16:15 - 17:45; ab 22.10.2009; INF 325 / SR 24; Frank, A.

Kommentar

Leistungsbewertung:

SS-CL, SS-FAL (Master): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

Inhalt

In diesem Seminar beleuchten wir vielfältige Phänomene der Semantik von Ereignissen.

Im Zentrum stehen dabei die diskurssemantischen Aspekte der Verbsemantik und ihre Relevanz für automatische Diskursverarbeitung und maschinelles Textverstehen.

Das Seminar behandelt zunächst die linguistischen Grundlagen sowie den Stand der Forschung zur Modellierung von Verbsemantik in computerlinguistischen Ressourcen, zu automatischen Verfahren für die Lexikonakquisition und zur automatischen Analyse von Ereignissen im Diskurs.

Die Semantik von Verben und Verbklassen steht in systematischer Beziehung zu diskurssemantischen Aspekten, die konstitutiv sind für das automatische Textverstehen. Wir betrachten insbesondere:

(i) implizite semantische Relationen zwischen Ereignissen bzw. Zuständen (z.B. Präsupposition, Implikation, Kausalität),

(ii) temporale Relationen zwischen Ereignissen und Zuständen im Diskurs, sowie deren Lokalisierung relativ zu Zeit und Raum,

(iii) die Interaktion von Verbsemantik und Diskursrelationen, wie sie vor allem in der SDRT im Vordergrund steht, bis hin zu:

(iv) Phänomenen der Anaphorik.

Neben der formal-linguistischen Analyse und Modellierung dieser Phänomene untersuchen wir vor allem datengetriebene Methoden für die Akquisition und die automatische Verarbeitung der Semantik von Ereignissen im Diskurs.

Leistungsnachweis Regelmäßige Teilnahme; aktive Mitarbeit; Referat und Hausarbeit oder Projekt

Vereinbarung von Referatsthemen: ab Oktober
Voraussetzung Programmierprüfung
Literatur Wird zu Beginn des Semesters bekanntgegeben

Unterspezifikationsformalismen für die semantische Verarbeitung - AS-CL, AS-FL, V01, SS-CL, SS-FAL

HpS; Nr.: 09-160-20-17; SWS: 2

Mo; wöch; 18:15 - 19:45; ab 12.10.2009; INF 325 / SR 24; Herweg, M.

Kommentar Leistungsbewertung:

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

SS-CL, SS-FAL (Master): 8 LP

Inhalt Der effiziente Umgang mit Mehrdeutigkeiten ist eine der größten Herausforderungen in der Sprachverarbeitung. Das Problem besteht darin, dass die natürliche Sprache voll von offensichtlichen oder versteckten Mehrdeutigkeiten ist und dass eine Grammatik umso mehr von diesen Mehrdeutigkeiten entdeckt, je umfassender sie die möglichen Konstruktionen einer Sprache abdeckt. Wenn nun für jede Lesart einer mehrdeutigen Konstruktion eigene Repräsentationen aufgebaut und als je eigene Analysepfade parallel oder nacheinander verfolgt werden müssen, stößt ein computerlinguistisches System schnell an seine Verarbeitungsgrenzen.

Aus diesem Grund wurde in den letzten Jahren eine Reihe von Verfahren entwickelt, die es erlauben, Mehrdeutigkeiten in einer kompakten Repräsentation darzustellen, die bezüglich der verschiedenen Lesarten unterspezifiziert ist. Erst wenn zusätzliche Information, z.B. aus dem sprachlichen Kontext oder aus der Äußerungssituation, bestimmte Lesarten ausschließt und andere favorisiert, werden die Repräsentationen sukzessive spezifischer.

Wir wollen uns in diesem Hauptseminar auf die wichtigsten Unterspezifikationsformalismen für die semantische Verarbeitung konzentrieren. Zunächst verschaffen wir uns einen Überblick über die einschlägigen linguistischen Phänomene und zentrale computerlinguistische Anwendungsbereiche für Unterspezifikation. Im Anschluss daran erarbeiten wir die wichtigsten Verfahren, die in computerlinguistischen Anwendungen eingesetzt werden.

Leistungsnachweis Referat und schriftliche Hausarbeit (Ausarbeitung des Referats)
Voraussetzung Logikkenntnisse (Einführung in die Logik), Grundkenntnisse in der Semantik

Die Sitzung am 12.10. wird für eine Auffrischung der Logik-Kenntnisse der TeilnehmerInnen verwendet (Schwerpunkt: modelltheoretische Semantik, Lambda-Kalkül).

Literatur Die Literatur für die Referate und Hausarbeiten wird zum Beginn des Seminars vorgestellt. Zur Vorbereitung werden die Kapitel über Semantik in:

* Carstensen, K.U., et al. (2001): Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg/Berlin: Spektrum Akademischer Verlag (darin Kap. 3.4)
* Görz, G., et al. (2000): Handbuch der Künstlichen Intelligenz. 3. Auflage, München/Wien: Oldenbourg Verlag (darin Kap. 19) empfohlen.

Die Teilnehmer/innen sollten sich darauf einstellen, dass der größte Teil der Literatur für Referate und Hausarbeiten nur in englischer Sprache vorliegt.

Paraphrasen und Inferenz - V01, SS-CL, SS-FAL, AS-CL, AS-FL

HpS; Nr.: 09-160-20-18; SWS: 2

Mi; wöch; 14:15 - 15:45; ab 21.10.2009; INF 327 / SR 3; Thater, S.

Kommentar Leistungsbewertung:

V01 (Bachelor, alte Prüfungsordnung): 6 LP

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

SS-CL, SS-FAL (Master): 8 LP

Inhalt Unter Paraphrasen versteht man sprachliche Ausdrücke, die annähernd dieselbe Bedeutung haben. Dass gleiche Information durch verschiedene sprachliche Ausdrücke realisiert werden kann, ist für die Qualität und Robustheit sprachtechnologischer Anwendungen häufig ein Problem. In diesem Seminar wollen wir verschiedene Techniken und Methoden für die automatische Identifikation und Generierung von Paraphrasen diskutieren (wobei wir von einem recht weit gefassten Paraphrasenbegriff ausgehen werden), sowie deren Anwendung und sprachverarbeitende Systeme.

Literatur wird in der ersten Sitzung bekanntgegeben

Algorithmische Linguistik

Maschinelle Übersetzung - CS-CL, BS-CL, BS-AC, A20

V; Nr.: 09-160-10-05; SWS: 2; LP: 4

Mo; Einzel; 09:15 - 12:45; 05.10.2009 - 05.10.2009; INF 366 / SR 12; Vorlesung; Eberle, K.

Mo; Einzel; 14:15 - 15:45; 05.10.2009 - 05.10.2009; INF 366 / SR 12; Vorlesung; Eberle, K.

Block; 09:15 - 12:45; 06.10.2009 - 09.10.2009; INF 327 / SR 4; Vorlesung; Eberle, K.

Block; 14:15 - 15:45; 06.10.2009 - 09.10.2009; INF 327 / SR 4; Vorlesung; Eberle, K.

Kommentar Leistungsbewertung:

CS-CL, BS-CL, BS-AC (Bachelor, neue Prüfungsordnung): 4 LP

A20 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt Nach einem kurzen Überblick über die Geschichte der Maschinellen Übersetzung werden die verschiedenen sog. regel-basierten Architekturen vorgestellt, die bis Ende der 90er Jahre die Maschinelle Übersetzung bestimmt haben (das sind vor allem die direkte Übersetzung, Transfer- und Interlingua-Verfahren). An Übersetzungsbeispielen und -schwierigkeiten werden die Vor- und Nachteile der Verfahren exemplifiziert.

Anhand der Entwicklungsumgebung des Übersetzungssystems translate wird Einblick in die Umsetzung von Spielarten der Transfer-Konzeption in einem kommerziellen System gegeben, insbesondere werden dabei Regeln aus verschiedenen System-Komponenten, wie lexikalischer Lookup, grammatische Analyse, Transfer und Generierung, exemplarisch skizziert und deren Wirkungsweise an Testbeispielen demonstriert.

Seit den 90er Jahren werden vermehrt andere, Korpus-basierte, Methoden für die Maschinelle Übersetzung diskutiert. Im zweiten Teil der Veranstaltung wird in solche Methoden, insbesondere die Grundlagen der sog. Statistik-basierten und der Beispiel-basierten Übersetzung eingeführt und am Beispiel von translate motiviert, wie Methoden kombiniert werden können.

Angesichts der zur Verfügung stehenden Zeit und der Vorkenntnisse ist das Lernziel, einen Eindruck zu vermitteln, über die Schwierigkeiten mit der eine Maschine bei der Übersetzung konfrontiert ist, über gegangene und mögliche Wege, die Aufgabe algorithmisch zu bewältigen und über die Vor- und Nachteile, die den verschiedenen Konzeptionen immanent sind.

Leistungsnachweis Klausur

Literatur **einführende Literatur** :

* Arnold, D., L. Balkan, R.L. Humphreys, S. Meijer & L. Sadler (1994): Machine Translation: An Introductory Guide, Oxford, NCC Blackwell. http:/www.essex.ac.uk/linguistics/clmt/MTbook/HTML/book.html
* Nirenburg, Sergei (ed.) (2003) Readings in Machine Translation.Cambridge: MIT Press.
* Schwanke, M. (1991): Maschinelle Übersetzung- Ein Überblick über Theorie und Praxis, Springer Verlag.
* Trujillo, A. (1999): Translation Engines: Techniques for Machine Translation, Springer Verlag.

weiterführende Literatur zu verschiedenen Methoden :

* Beaven, J. (1992): Shake and Bake Machine Translation, in COLING92.
* P. F. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, & P. Roossin, "A Statistical Approach to Machine Translation," Computational Linguistics 16(2), 1990.
* Carl, M., Way, A. (ed.) (2003): Recent Advances in Example-Based Machine Translation, Kluwer Academic Publishers, Dordrecht.
* Manning, Christopher D., Schütze, Hinrich: Chap. 13 Statistical Alignment and Machine Translation. In: Manning, Schütze: Foundations of Statistical NLP, 1999
* Michael McCord: Design of LMT, in: Computational Linguistics (15) 1989
* Sumita, E. , Iida, H., Kohyama, H.: Translating with Examples. In: A New Approach to Machine Translation. The Third International Conference on Theoretical and Methodological Issues in Machine Translation, 1990.

Parsing - ACL, B09

V/Ü; Nr.: 09-160-08-01; SWS: 2

Di; wöch; 11:15 - 12:45; ab 20.10.2009; INF 327 / SR 1; Thater, S.

Kommentar	Leistungsbewertung: ACL (Bachelor, neue Prüfungsordnung): 6 LP B09 (Bachelor, alte Prüfungsordnung): 4 LP
Inhalt	Die Vorlesung stellt verschiedene Verfahren für die Syntaxanalyse (Parsing) vor. Neben klassischen Strategien (top-down, bottom-up, left-corner) und Algorithmen für kontextfreie Grammatiken werden wir Parsing-Verfahren für lexikalisierte und unifikationsbasierte Grammatikformalismen sowie stochastische Erweiterungen bestehender Ansätze betrachten.
Leistungsnachweis	Klausur und erfolgreiche Bearbeitung der Übungsaufgaben
Voraussetzung	Formale Grundlagen, Einführung in die Computerlinguistik
Literatur	* Naumann und Langer: Parsing. B. G. Teubner, 1994. * Grune und Jacobs: Parsing Techniques. 2. Auflage. Springer, 2008. * Stuart M. Shieber, Yves Schabes & Fernando C. N. Pereira (1995). Principles and implementation of deductive parsing. The Journal of Logic Programming Volume 24, Issues 1-2. Weitere Literatur wird zu Beginn der Veranstaltung bekanntgegeben.

Einführung in die statistische Sprachverarbeitung - FF-SM, A10

V/Ü; Nr.: 09-160-09-01; SWS: 2

Mi; wöch; 14:15 - 15:45; ab 21.10.2009; INF 306 / SR 13; Ponzetto, S.

Kommentar	Leistungsbewertung: FF-SM (Bachelor, neue Prüfungsordnung): 6 LP A10 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt

Statistische NLP-Methoden sind de-facto der Standardansatz in der aktuellen NLP-Forschung. Dieser Kurs wird eine Einführung in die theoretischen sowie in die praktischen Grundlagen der Statistischen NLP geben.

Der Schwerpunkt des Kurses wird data-driven sein, d.h. die Studierenden werden mit großen Korpora arbeiten und sie werden lernen, große Datenmengen zu handhaben. Die Anwendung von statistischen NLP-Methoden wird uns z.B. ermöglichen, Kollokationen und N-Gramme zu analysieren und diese für Textkategorisierung zu verwenden.

Wir werden uns mit einer Auswahl von bestimmten NLP-Anwendungen befassen, z.B. PoS-Tagging und Parsing, obwohl diese Methoden auf eine Vielzahl anderer NLP-Themen übertragbar sind. Als solches bietet der Kurs eine Grundlage für fortgeschrittene NLP-Themen, z.B. Maschinelle Übersetzung.

Von den Studierenden wird erwartet, dass sie ein gutes Verständnis der Theorie entwickeln und in der Lage sind, einfache NLP-Anwendungen, wie z.B. ein Hidden Markov Model oder einen Maximum Entropy basierten PoS-Tagger, zu implementieren.

Leistungsnachweis

Wöchentliche Hausaufgaben (Übungen sowie Programmieraufgaben)

Schriftliche Abschlussklausur

Zur Klausur wird nur zugelassen, wer mindestens 80% der Übungsaufgaben bearbeitet hat und mindestens 60% der maximalen Punktzahl erreicht hat.

Voraussetzung

Voraussetzung ist der erfolgreiche Abschluss der Kurse "Einführung in die Computerlinguistik" sowie "Formale Grundlagen". Programmierkenntnisse (auf dem Niveau von Programmieren I) sind für die Lösung der Übungsaufgaben von Vorteil.

Literatur

* Daniel Jurafsky and James H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Second Edition. Prentice Hall.
* Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press.
* Natural Language Toolkit: http://nltk.sourceforge.net/index.php/Book

Information Retrieval - V01, SS-TAC, SS-CL, AS-CL

HpS; Nr.: 09-160-20-07; SWS: 2

Mo; wöch; 11:15 - 12:45; ab 19.10.2009; INF 325 / SR 24; Haenelt, K.

Kommentar

Leistungsbewertung:

AS-CL (Bachelor, neue Prüfungsordnung): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

SS-CL, SS-TAC (Master): 8 LP

Inhalt

Information Retrieval Systeme sollen Informationssuchende dabei unterstützen, aus großen elektronisch verfügbaren Informationsmengen (Texte, Datenbanken, multimediale Dokumente) passende Information herauszufinden. Im Seminar sollen die verschiedenen Ansätze und grundlegende Methoden und Algorithmen solcher Systeme erarbeitet und vermittelt werden.

Leistungsnachweis

Durchführung eines Seminarprojektes und ein Referat

Voraussetzung

Zwischenprüfung in Computerlinguistik oder vergleichbare Kenntnisse, Programmierkenntnisse (möglichst C/C++/JAVA), Programmierprüfung

Einführung in die Computerlinguistik - ICL, B01

V/Ü; Nr.: 09-160-01-01; SWS: 4; LP: 6

Di; wöch; 09:15 - 10:45; ab 13.10.2009; INF 350 / OMZ R U013; Frank, A.

Do; wöch; 11:15 - 12:45; ab 22.10.2009; INF 350 / OMZ R U013; Frank, A.

Kommentar

Leistungsbewertung:

ICL (Bachelor, neue Prüfungsordnung): 6 LP

B01 (Bachelor, alte Prüfungsordnung): 6 LP

Inhalt

Die Vorlesung führt ein in die Grundlagen, zentralen Fragestellungen und Methoden der Computerlinguistik. In einem Gesamtüberblick werden die wesentlichen Grundlagen der Computerlinguistik eingeführt:

* Ebenen der Sprachbeschreibung (Phonologie, Morphologie, Syntax, Semantik, Pragmatik),
* formale mathematische und logische Modelle zur Beschreibung der entsprechenden linguistischen Phänomene und
* algorithmische Verfahren zur automatischen Verarbeitung auf Basis dieser Modelle.

Dabei nähern wir uns speziellen Problemen und Fragestellungen der Computerlinguistik und ihren spezifischen Lösungsstrategien. Spezielle Themen werden sein: Ambiguitätsbehandlung, Approximierung sprachlicher Regularitäten, syntaktische und semantische Verarbeitung.

Die Vorlesung gibt einen Überblick über computerlinguistische Anwendungen, diskutiert das Verhältnis zu Nachbardisziplinen, und führt durch praktische Übungen in die speziellen Fragestellungen einzelner Teilgebiete der Computerlinguistik ein.

Leistungsnachweis

* Erfolgreiche Bearbeitung der Übungsaufgaben (mind. 60%)
* Erfolgreich bestandene Klausur
* Aktive Teilnahme

Regelmäßige Präsenz ist Voraussetzung für den Scheinerwerb.

Die erfolgreich bestandene Klausur ist Teil der Orientierungsprüfung.

Literatur

* Daniel Jurafsky and James H. Martin (2000): Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall Series in Artificial Intelligence. Prentice Hall.
* Kai-Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde, Hagen Langer (Hrsg.) (2004): Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg: Spektrum, Akademischer Verlag.
* Natural Language Toolkit, NLTK: http://nltk.sourceforge.net/index.php/Book

Grundlagen Semantic Web - CS-CL, A05

V; Nr.: 09-160-10-10; SWS: 2; LP: 4

Mo; Einzel; 09:15 - 12:45; 05.10.2009 - 05.10.2009; INF 306 / SR 14; Vorlesung; Rudolph, S.

Block; 09:15 - 12:45; 06.10.2009 - 09.10.2009; INF 306 / SR 14; Vorlesung; Rudolph, S.

Block; 14:15 - 16:45; 06.10.2009 - 09.10.2009; INF 306 / SR 14; Vorlesung; Rudolph, S.

Kommentar

Leistungsbewertung:

CS-CL, BS-CL, BS-AC (Bachelor, neue Prüfungsordnung): 4 LP

A05 (Bachelor, alte Prüfungsordnung): 4 LP

Inhalt

Der Begriff Semantic Web bezeichnet allgemein eine Erweiterung des World Wide Web durch Metadaten und Anwendungen mit dem Ziel, die Bedeutung (Semantik) von Daten im Web für intelligente Systeme z.B. im E-Commerce und in Internetportalen nutzbar zu machen. Eine zentrale Rolle spielen dabei die Repräsentation und Verarbeitung von Wissen in Form von Ontologien. In dieser Vorlesung werden die Grundlagen der Wissensrepäsentation und -verarbeitung für die entsprechenden Technologien vermittelt sowie Anwendungsbeispiele vorgestellt. Dabei werden folgende Themenbereiche betrachtet:

* Grundlagen von XML (Extensible Markup Language) und XML Schema
* RDF (Resource Description Framework) und RDF Schema zur Darstellung von Metadaten und einfachen Ontologien
* Die Web Ontology Language (OWL) und ihre aktuelle Erweiterung OWL 2
* Die SPARQL-Anfragesprache für RDF, konjunktive Anfragen für OWL
* Regelsprachen für das Semantic Web

	* Praktische Anwendungen
Leistungsnachweis	Leistungsnachweis durch Klausur
Literatur	Literatur wird im Kurs bekannt gegeben.

Spracherkennung - CS-CL, BS-CL, BS-AC, A18

PS; Nr.: 09-160-10-12; SWS: 2

Fr; 14täg.; 10:15 - 13:45; ab 23.10.2009; INF 325 / PCPool; Günther, C.;Klehr, M.

Kommentar	Leistungsbewertung:
	CS-CL, BS-CL, BS-AC (Bachelor, neue Prüfungsordnung): 4 LP
	A18 (Bachelor, alte Prüfungsordnung): 4 LP
Inhalt	Der Kurs wird die Grundlagen der Spracherkennung behandeln. Es werden die verschiedenen Verarbeitungsschritte der automatischen Spracherkennung behandelt: von der Signalverarbeitung bis zum Sprachmodell. Dabei wird auf aktuelle Forschungen auf diesem Gebiet eingegangen. Aber auch aktuelle Implementationen und Systeme (wie der IBM WebSphere Voice Server) sollen vorgestellt werden. Im praktischen Teil des Seminars wird auf der Grundlage von VoiceXML ein Sprachdialogsystem implementiert. Es werden die einzelnen Schritte des Entwurfs und der Implementierung behandelt (Wizard-of-Oz Test, Dialogmodell, Grammatikentwurf, Prompt-Design, Test). Es werden dabei die verschiedenen Einflussfaktoren wie Vokabulargröße oder Grammatikkomplexität auf das Erkennungsergebnis untersucht.
Leistungsnachweis	Ausarbeitung einer Programmieraufgabe (Sprachdialog-Modul)
Voraussetzung	Kenntnisse in Statistik und Signalverarbeitung sind von Vorteil, aber nicht erforderlich. Im Kurs werden Übungsaufgaben in VoiceXML gelöst, so dass Programmiererfahrungen (Java Script, XML) ebenfalls von Vorteil sind.
Literatur	* C. Günther, M. Klehr: VoiceXML 2.0, mitp 2003 * F. Jelinek: Statistical Methods for Speech Recognition, MIT Press 1997 * E. G. Schukat-Talamazzini: Automatische Spracherkennung, Vieweg 1995 * B. Eppinger, E. Herter: Sprachverarbeitung, Hanser 1993

Natural Language Generation for Virtual Environments - CS-CL, BS-CL, AC, A13

PS; Nr.: 09-160-10-22; SWS: 2

Mo; wöch; 16:15 - 17:45; ab 19.10.2009; INF 325 / SR 24; Roth, M.

Kommentar	Leistungsbewertung:
	CS-CL, BS-CL, AC (Bachelor, neue Prüfungsordnung): 6 LP
	A13 (Bachelor, alte Prüfungsordnung): 4 LP
Inhalt	Sprachgenerierung (auch Natural Language Generation, kurz NLG, genannt) bezeichnet ein Teilgebiet der Computerlinguistik, das sich mit der sprachlichen Realisierung aus semantischen/logischen Repräsentationen befasst. Dabei kann diese Aufgabe als komplexer Prozess verstanden werden, der aus verschiedenen Teilaufgaben wie beispielsweise Inhalts- und Diskursplanung, Wortwahl und Oberflächenrealisierung besteht. Dieser Kurs gibt eine Einführung in die Sprachgenerierung mit dem Ziel ein eigenes Generierungssystem zu planen und zu implementieren. In den ersten Wochen des Kurses werden wir ausgewählte Publikationen zur Sprachgenerierung aufarbeiten und diskutieren, um eine Grundlage für den zweiten Kursteil zu legen. Im zweiten Teil wollen wir dann die gewonnenen Einsichten anwenden und in Gruppenarbeit ein System entwickeln, welches sprachliche Anweisungen in virtuellen Umgebungen generieren soll. Durch die Mitarbeit im Kurs bietet sich die Möglichkeit zur Teilnahme an der GIVE-Challenge (http://www.give-challenge.org/), einem international organisierten Wettbewerb von Sprachgenerierungssystemen.
Leistungsnachweis	* Lektüre der zugrundegelegten Literatur * Aktive und regelmäßige Teilnahme

* Gruppenprojekt und schriftliche Ausarbeitung
* Je nach Teilnehmerzahl ggf. Referat

Voraussetzung Programmierprüfung

Machine Learning - SS-CL, SS-TAC

HpS/Ü; Nr.: 09-160-20-03; SWS: 3

Do; wöch; 11:15 - 12:00; ab 22.10.2009; INF 325 / SR 24; Fendrich, S.

Do; wöch; 14:15 - 15:45; ab 22.10.2009; INF 325 / SR 24; Fendrich, S.

Kommentar Leistungsbewertung:

SS-CL, SS-TAC (Master): 8 LP

Inhalt Die Veranstaltung hat in der ersten Semesterhälfte die Form einer Vorlesung; in der zweiten Hälfte erfolgen Referate der Teilnehmer. Gegenstand der Veranstaltung sind grundlegende Methoden des Maschinellen Lernens. Behandelt werden u.a. Entscheidungsbäume, Clustering-Verfahren, Bayessches Lernen, Kernel-basierte Methoden und Support-Vector-Maschinen. Darüber hinaus wird es im Rahmen einer Übung die Gelegenheit geben, die Data-Mining-Software WEKA kennen zu lernen.

Leistungsnachweis - Bearbeitung der Übungsaufgaben

- Referat (40%)
- mündliche Prüfung (60%)

Voraussetzung - Programmierprüfung

- Formale Grundlagen oder Mathematischer Vorkurs
- Grundkenntnisse in Statistik

Literatur
* Mitchell: Machine Learning. McGraw-Hill, 1997.
* Bishop: Pattern Recognition and Machine Learning. Springer, 2006.
* Witten/Frank: Data Mining. Morgan Kaufman, 2005.

weitere Literatur wird im Kurs bekannt gegeben

Data-Driven Grammar Induction - V01, AS-CL, AS-FL, SS-CL, SS-FAL

HpS; Nr.: 09-160-20-06; SWS: 2

Mi; wöch; 11:15 - 12:45; ab 21.10.2009; INF 327 / SR 1; Frank, A.

Kommentar Leistungsbewertung:

SS-CL, SS-FAL (Master): 8 LP

V01 (Bachelor, alte Prüfungsordnung): 6 LP

AS-CL, AS-FL (Bachelor, neue Prüfungsordnung): 8 LP

Inhalt Seit den 80/90er Jahren wurden linguistisch motivierte und formal wohldefinierte Grammatikformalismen entwickelt, insbesondere Lexical-Functional Grammar (LFG), Combinatory Categorial Grammar (CCG), Head-driven Phrase-Structure Grammar (HPSG) und Lexicalised Tree-Adjoining Grammar(LTAG). Durch die Entwicklung effizienter Parsingalgorithmen ist der Einsatz dieser Grammatikformalismen in computerlinguistischen Anwendungen realistisch geworden. Die Entwicklung umfangreicher manuell definierter Grammatiken ist zeitaufwendig und teuer; für multilinguale Sprachverarbeitung müssen jedoch umfangreiche und robuste Grammatiken in kurzer Zeit entwickelt werden.

Das Seminar führt ein in die Methodik der automatischen Induktion probabilistischer Grammatiken aus Baumbanken am Beispiel von PCFGs. Wir diskutieren insbesondere spezielle Verfahren für die automatische Induktion lexikalisierter und constraint-basierter Grammatiken (wie LFG, TAG, CCG und HPSG) aus angereicherten Baumbanken bzw. Baumbankgrammatiken. Hierbei werden wir die Charakteristiken der jeweiligen Grammatikformalismen und die entsprechenden Unterschiede der entsprechenden Grammatikinduktionsverfahren herausarbeiten. Abschließend widmen wir uns neueren Ansätzen für die Grammatikinduktion auf Basis paralleler Korpora.

Leistungsnachweis Lektüre der zugrundegelegten Literatur, Referat und Hausarbeit oder Referat und Projekt

Voraussetzung Programmierprüfung, Kenntnisse in Syntax

Literatur
* Aoife Cahill (2008): Treebank-Based Probabilistic Phrase Structure Parsing in: Language and Linguistics Compass 2/1, Blackwell, pp. 36-58.
* Daniel Jurafsky and James Martin (2008): Speech and Language Processing, Kap. 13 und Kap. 14
* ESSLLI Course 2006: Josef van Genabith, Julia Hockenmaier and Yusuke Miyao: Treebank-Based Acquisition of LFG, HPSG and CCG Resources

Weitere Literatur wird zu Beginn des Semesters bekanntgegeben.

Word Sense Disambiguation - V01, SS-CL

HpS; Nr.: 09-160-20-19; SWS: 2; LP: 8

Di; wöch; 09:15 - 10:45; ab 20.10.2009; INF 325 / SR 24; Ponzetto, S.

Inhalt Word Sense Disambiguation (WSD) is the problem of identifying the intended meaning (or sense) of a word, based on the context in which it occurs. Correctly identifying the senses of words in context is a central problem for Natural Language Processing (NLP), and robust performance on this task is accordingly expected to provide crucial lexical semantic information for many NLP applications such as machine translation, information retrieval, etc.

This seminar will provide a gentle introduction to state-of-the-art approaches in WSD. These include:

* knowledge-based methods that either (a) make use of dictionaries and thesauri and/or (b) manually crafted graph-like resources such as e.g. WordNet or GermaNet;
* supervised machine learning methods that learn classifiers from sense annotated data;
* minimally supervised methods (aka bootstrapping) that, starting with a small amounts of labeled data (seeds), iteratively harvest new sense annotations to improve the sense disambiguation accuracy.

Students will present current work from the literature in short, seminar-format presentations (Referate). In addition, every 4-6 weeks they will be expected to form small groups of 3-4 people and work on a project, e.g. implement and/or extend an existing state-of-the-art WSD approach. Each of the groups is expected to submit a short report (2-4 pages), as well as to give a short project-overview presentation at the end of each round. Students are expected to *actively* participate in the class discussions during their fellow students' presentations, as well as in the seminar's projects. This means that you'll have to read the papers **before** the class period in which they will be presented and discussed, as well as **clearly** present to the audience what your specific work was as part of the seminar's projects.

Determination of final grade:

33%: presentation

33%: participation in the seminar's projects

33%: participation in the class discussions

Leistungsnachweis Aktive Teilnahme und regelmäßige Abgabe von Projektarbeit in kleinen Gruppen. Vortrag/Präsentation.

Voraussetzung Voraussetzungen sind die bestandene Zwischenprüfung (Magister) und Programmierprüfung. Vorkenntnisse in statistischer NLP oder Maschinellem Lernen sind von Vorteil.

Vollständige Lektüre von Navigli (2009) (siehe unten)

Literatur
* R. Navigli. Word Sense Disambiguation: a Survey, ACM Computing Surveys, 41(2), ACM Press, 2009, pp. 1-69 (WICHTIG: Lektüre zu Semesterbeginn vorausgesetzt!); Link: http://www.dsi.uniroma1.it/~navigli/pubs/ACM_Survey_2009_Navigli.pdf

* Eneko Agirre & Philip Edmonds (eds.) Word Sense Disambiguation Algorithms and Applications, Springer, 2006 (wird als Referenz benutzt) Link: http://www.wsdbook.org/