NLP - Analyse des Wissensrohstoffs Text

Dr. Andreas Hotho
Dominik Benz
Beate Krause

Sommersemester 2008

Organisatorisches

Vorlesung

- Beginn: 8. April 2008
- Dienstag 10.15 h - 11.45 h, in Raum 1607 oder 0443

Übungen

- Beginn: 16. April 2008
- Mittwochs, 10.15 h - 11.45 h, in Raum 1418 (Altbau WA 73)
- wird als Präsenz- und Praxisübung abgehalten (s. nächste Folie)
- Programmierhausaufgaben

Unterlagen

- siehe Literatur

Prüfung

- Die Prüfung wird je nach Teilnehmerzahl mündlich oder schriftlich abgehalten.

Organisatorisches

- Mailingliste für alle Studenten die am Fachgebiet eine Vorlesung hören.
- Die Mailingliste hat den Namen „kde-stud“
- Um sich einzutragen gehen sie bitte auf die folgende Webseite: https://mail.cs.uni-kassel.de/mailman/listinfo/kde-stud
- Die E-Mail-Adresse der Liste lautet: kde-stud@cs.uni-kassel.de
- Über diese Liste werden wir zusätzliche Ankündigungen und Informationen schicken
- Sie können dort auch Fragen stellen oder diskutieren

Organisatorisches

Präsenzübung bedeutet

- **selbständiges Bearbeiten** des Übungsblattes in Kleingruppen à 3-4 Personen unter Betreuung des Assistenten
- **kein prinzipielles Wiederholen** des Vorlesungsstoffs
- **kein Vorrechnen** der Musterlösung etc.
  (Diese wird später zur Verfügung gestellt.)
- **Nötig dafür:**
  - selbständige Vorlesungsnachbereitung **vor** der Übung
  - Mitbringen des Skriptes
  - eigene Aktivität entfalten

Organisatorisches

Warum ein neues Übungskonzept?

- aktives Erarbeiten des Vorlesungsstoffes bringt mehr
- Zusammenhänge im Stoff erkennen
- strukturiertes Denken und selbständiges Arbeiten lernen
- Teamarbeit lernen
- Erklären lernen (als Tutor und als Teilnehmer)
- Klausurtraining ;-)
- *Ihr Studium der ... haben Sie abgeschlossen. Zu Ihren persönlichen Stärken zählen Sie Eigeninitiave, Kommunikations- und Kooperationsbereitschaft, Teamarbeit.* (Typischer Anzeigentext)

Organisatorisches

Praxisübung – Implementieren einer Suchmaschine

- Ausgabe der ersten Praxisaufgabe zur ersten Übung am 16.4.2008
- Am 23.4.2008 Fragestunde zur Praxisaufgabe
- Abgabe der ersten Praxisaufgabe bis 29.4.2008 12:00 per Email
- Präsentation des Ergebnisses am folgenden Tag
- Praxisaufgaben im 14 Tagesrhythmus
- 4 von 5 Aufgaben müssen für einen Notenbonus von 0.3 abgegeben werden

Organisatorisches

Sprechstunden nach Absprache:

Andreas Hotho:	hotho@cs.uni-kassel.de	0561/804-6252
Dominik Benz:	benz@cs.uni-kassel.de	0561/804/6266
Beate Krause:	krause@cs.uni-kassel.de	0561/804/6254

FG Wissensverarbeitung, FB Mathematik/Informatik
Raum 0440, Wilhelmshöher Allee 73

Informationen im Internet: http://www.kde.cs.uni-kassel.de

Hier ist u.a. folgendes zu finden:

- aktuelle Ankündigungen
- Folienkopien
- Übungsblätter
- Literaturempfehlungen
- Termine

ENDOWED CHAIR OF THE HERTIE FOUNDATION
Knowledge and Data Engineering
DEPARTMENT OF MATHEMATICS & COMPUTER SCIENCE

Empfohlene Literatur

- Manning Ch. D./H. Schütze (1999). Foundations of Statistical Natural Language Processing. Cambridge: The MIT Press.
- Carstensen, K./Ch. Ebert/C. Endriss/S. Jekat/R. Klabunde/H. Langer (2004). Computerlinguistik und Sprachtechnologie. Eine Einführung. 2. Aufl. Heidelberg: Elsevier.
- Jurafsky, D./J.H. Martin (2000). Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River: Prentice Hall.

Übersicht

- Einführung
- Eigenschaften von Text
- Words I: Satzgrenzenerkennung, Tokenization, Kollokationen
- Words II: N-Gram-Modelle, Morphologie
- Tagging I: Transformationsbasiertes Tagging
- Tagging II: Hidden Markov Modelle
- Parsing I: Kontextfreie Grammatiken (CFG)
- Parsing II: probabilistisches Parsing
- Semantik I: Lexikalische Semantik (Lexeme, Homonymie, Homographie, Homophonie...)
- Semantik II: Wortbedeutungsdisambiguierung
- Applikationen I: Text Summarization
- Applikationen II: Textübersetzung, Wortsinnerkennung...

Viele der Vorlesungsfolien wurden aus der Vorlesung:

„Symbolische und statistische Verfahren“

übernommen.

Jan Strunk

Gegenstand der Computerlinguistik

- Gegenstand der Computerlinguistik sind Formalismen, Algorithmen und Verfahren zur maschinellen Verarbeitung natürlicher Sprache.
- Es geht also darum, wie man mit dem Computer natürliche Sprache verarbeiten kann.
- Sind Ihnen schon Beispiele für Sprachtechnologie begegnet?
  - Maschinelle Übersetzung (z.B. bei Google oder Alta Vista)
  - Spracherkennung (Diktiersoftware, Sprachsteuerung)
  - Rechtschreib- und Grammatikprüfung
  - Einfache Dialogsysteme z.B. beim Telefonbanking
  - …

Die (weitentfernten) Ziele der Computerlinguistik

- Kommunikation mit dem Computer in natürlicher Sprache (Sprachverstehen)
  - Beispiele: Star Trek, HAL, etc.
- Bearbeitung sprachlicher Aufgaben durch den Computer
  - Automatische Generierung von Texten
  - Übersetzung
  - Korrektur
  - Usw.
- Erschließung von Wissen aus Texten
  - Semantisches Web
  - Informationsextraktion

Definition Natural Language Processing (NLP)

- „Verarbeitung Natürlicher Sprache“ (VNS)
  - ist ein Teilbereich der Computerlinguistik
  - befasst sich mit
    - dem automatischen **Erzeugen** sowie
    - dem automatischen **Verstehen**

  natürlicher (menschlicher) Sprache (in geschriebener / gesprochener Form)

- Teilbereiche:
  - Formalismen zur Repräsentation der Bedeutung
  - Algorithmen zur „Transformation“ Bedeutung ⇔ Sprache
    - Symbolische / Statistische Ansätze (→ „Statistical NLP“, s.u.)

Gegenstand der Computerlinguistik: Beispiel MÜ

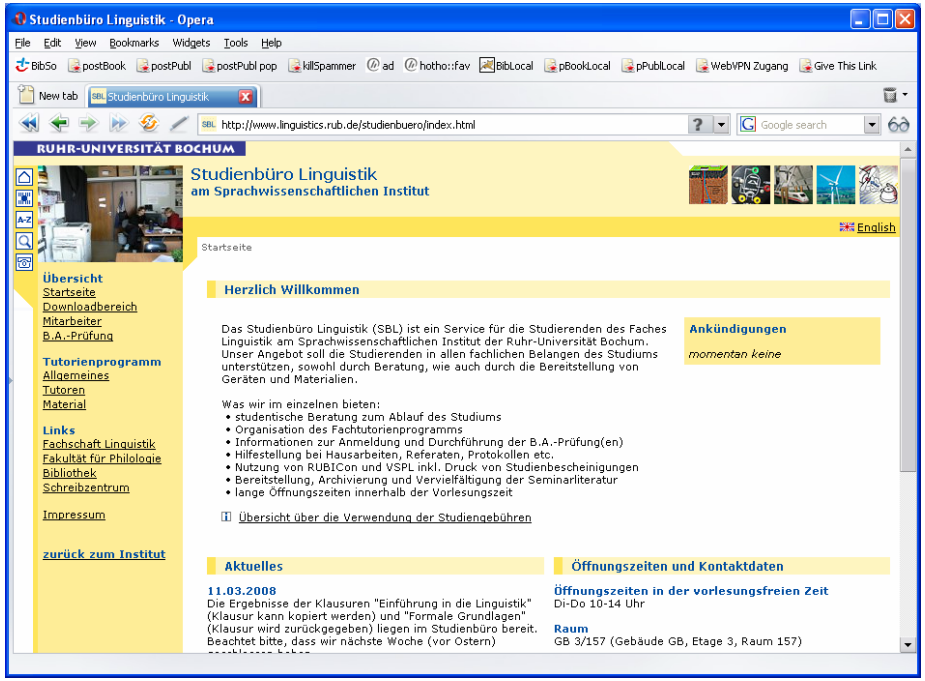
- Was sind Teilaufgaben bei der maschinellen Übersetzung?
  - Analyse der Struktur des Satzes in der Ursprungssprache
  - Analyse der Bedeutung in der Ursprungssprache
  - Übertragung der Bedeutung in die Zielsprache
    - Auswahl geeigneter Lexeme und Konstruktionen
  - Generierung einer grammatischen Struktur in der Zielsprache

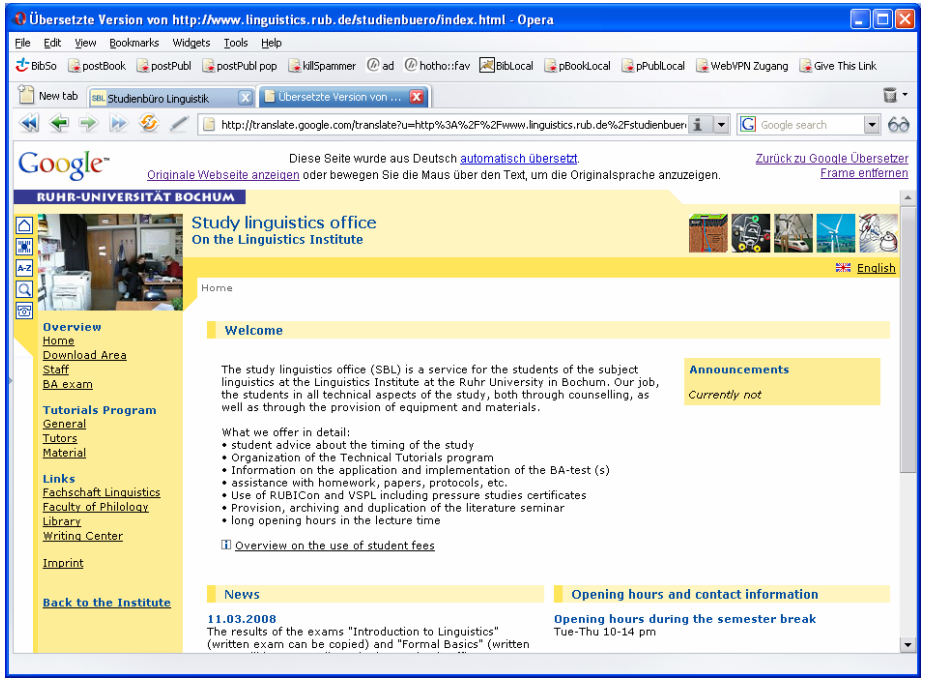
- Besondere Herausforderungen, die die automatische Sprachverarbeitung schwierig machen
  - Ambiguität (Mehrdeutigkeit)
    - Lexikalisch oder strukturell
  - Produktivität (unbegrenzte Anzahl von möglichen Sätzen und Wörtern)
  - Kontextabhängigkeit von Bedeutung und Form

Gegenstand der Computerlinguistik: Beispiel MÜ

- Beispiel:
  - Lassen Sie von Google die Seite des Studienbüros der Linguistik vom Deutschen ins Englische übersetzen
  - http://www.linguistics.rub.de/studienbuero/index.html
  - http://translate.google.com/translate_t?hl=de

  - Auftretende Probleme:
    - Unvollständige oder falsche syntaktische Analyse
    - Lexikalische Ambiguität (Mehrdeutigkeit)
      Ablauf = *Verlauf* und *Ablauf* = *Ende*
    - Unbekannte Wörter (z.B. *Studienbüro*)
    - Korrekte Interpretation des Pronomens *ihr* abhängig vom Kontext
      *Ab sofort findet **ihr** auf unserer Seite auch Infos zur B.A.-Prüfung.*
      *From now finds **her** on our side also about the B.A.-Prüfung.*





Gegenstand des Kurses

- Wir werden uns in diesem Kurs allerdings nicht intensiv mit solch komplexen Problemen wie der automatischen Übersetzung oder der Steuerung des Computers mittels natürlicher Sprache beschäftigen
- Beschränkung auf die Verarbeitung geschriebener Sprache
- Betrachtung von grundlegenden Problemen bei der Verarbeitung natürlicher Sprache
  - Auflösung von Ambiguität (Disambiguierung)
  - Umgang mit Produktivität
  - Modellierung von Kontextabhängigkeit

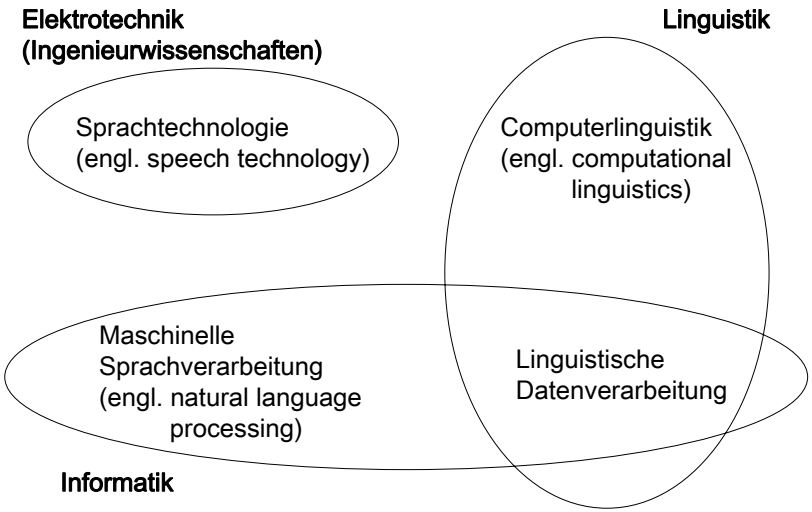
Gegenstand des Kurses

- Vorstellung grundlegender Algorithmen und Ansätze zur Lösung dieser Probleme
  - Vorverarbeitung von Textdaten
  - Strukturelle Analyse (Parsing)
  - Klassifikation (z.B. zur Disambiguierung)
  - Maschinenlernverfahren
- Verfahren zum Testen und zur Evaluation von computerlinguistischen Systemen

Gegenstand des Kurses

- Am Ende des Kurses sollten Sie also gelernt haben
  - Welche Herausforderungen bei der Verarbeitung natürlicher Sprache auftreten,
  - Was gängige Ansätze zur Lösung dieser Herausforderungen sind (illustriert an Hand einzelner Teilprobleme),
  - Was es für Standards und Verfahren bei der Evaluation computerlinguistischer Systeme gibt.
  - Einblicke in einige der typischen Anwendungsfelder der Computerlinguistik bekommen haben.
- Sie sollten dann auch fähig sein, kleinere computerlinguistische Systeme selbständig zu implementieren und zu evaluieren.

Beschäftigung mit der Verarbeitung menschlicher Sprache



Computerlinguistik und ihre Nachbardisziplinen

Elektrotechnik
- Signalverarbeitung
- Sprachtechnologie als Ingenieurwissenschaft

Mathematik
- Komplexitätstheorie
- Graphentheorie
- Logik
- Statistik

Computerlinguistik
- Formalismen, Algorithmen und Verfahren zur Verarbeitung natürlicher Sprache (Theorie)
- Praktische Systeme

Informatik
- Komplexitätstheorie
- Formale Sprachen
- Automaten
- Suchalgorithmen
- Maschinenlernen
- Information Retrieval

Linguistik (Sprachwissenschaft)
- Theorien zur Struktur und Komplexität menschlicher Sprache
- Analyse einzelner Sprachen
- Formalismen

Künstliche Intelligenz (KI)
- Wissensrepräsentation
- Maschinenlernen
- Planungsverfahren

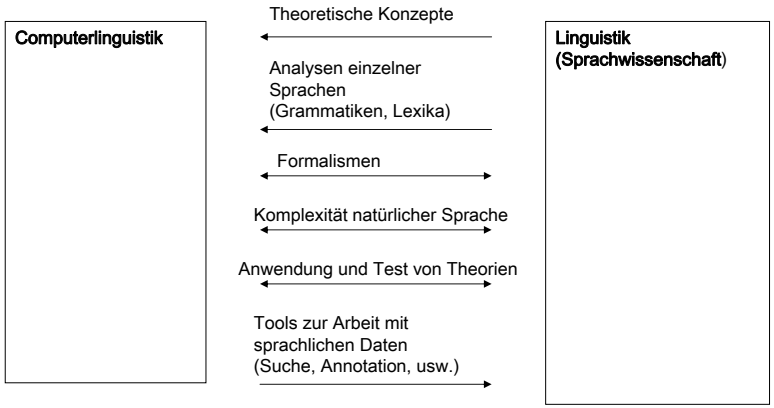
Philosophie
- Sprachphilosophie
- Logik
- Bedeutungstheorien

Psychologie / Cognitive Science
- menschliche Sprachverarbeitung
- Sprachlernen beim Menschen

(Frei nach Klabunde et al. 2004)

Computerlinguistik und ihre Nachbardisziplinen

- Verhältnis von Computerlinguistik und theoretischer Linguistik



Symbolische und statistische Verfahren

- Symbolische Verfahren
  - Basieren auf Regeln
  - Was sind mögliche / unmögliche Strukturen?
  - Regeln werden meist von menschlichen Experten formuliert
  - Beispiel: Strukturanalyse (Parsing) eines Satzes mit Hilfe einer formalen Grammatik
- Statistische (stochastische) Verfahren
  - Basieren auf statistischen Modellen, die auf einer großen Menge von Daten trainiert werden („datengetrieben“)
  - Was sind wahrscheinliche / unwahrscheinliche Strukturen?
  - Daten werden oft von menschlichen Experten annotiert
  - Beispiel: Sprachmodelle – z.B. im Handy
    „Welches Wort ist am wahrscheinlichsten gegeben eine Folge von mehrdeutigen Eingaben und den vorangegangenen Kontext?“

Symbolische und statistische Verfahren – Geschichtlicher Abriss

- Erste Überlegungen zur MÜ auf Basis der Informationstheorie (Stochastik)
- Chomskys (1957) Behauptung, dass statistische Ansätze nicht der Produktivität der Sprache gerecht werden:
  Colorless green ideas sleep furiously. vs.
  **Furiously sleep ideas green colorless.*
  (Nach Chomskys Ansicht beide gleich unwahrscheinlich)
- Symbolische Verfahren in der syntaktischen und semantischen Analyse und in der maschinellen Übersetzung
- Statistische Verfahren bei der Erkennung gesprochener Sprache
- Heute: Kombination symbolischer und statistischer Verfahren, um die Vorteile beider Paradigmen zu kombinieren
  - Z.B. Probabilistische kontextfreie Grammatiken
  - Vermehrte Einbindung von linguistischem Wissen in Sprachmodelle zur automatischen Spracherkennung

Geschichte der Computerlinguistik

Abbildung 1.1: Zeitliche Entwicklung der Computerlinguistik-Konzepte (links) und Anwendungen (rechts)

(aus Klabunde et al. 2004, S. 22)

Heutige Bedeutung der Computerlinguistik

- Handys
  - Texteingabe (T9)
- Sprachausgabe
  - Navigationsgeräte
- Sprachsteuerung
  - Navigationsgeräte
  - Handys
  - Dialogsysteme (Telefonservice)
- Internet
  - Texttechnologie
  - Informationsextraktion
  - Information Retrieval
  - Übersetzung
- Semantic Web
  - Suche nach Konzepten statt nach Wörtern
  - Automatische Informationserschließung