Vorbesprechung Mathe III

Dr. Vera Demberg, Prof. Dr. Enrico Lieblang (HTW)

Universität des Saarlandes

April 19th, 2012

Formalien

- Pflichtveranstaltung im Bachelor-Studiengang Computerlinguistik, vorgesehen für das 2. Semester (9 LP)
- Mo & Di 16–18h, Do 14–16h (immer c.t.)
  2x Vorlesung, 1x Übung (flexibel)
- Ort: Seminarraum der Computerlinguistik oder CIP-Raum
- Erste Hälfte: Prof. Lieblang (HTW)
- Zweite Hälfte: Dr. Demberg
- Literatur z.T. auf Englisch
  - Manning & Schütze (1999):
    Foundations of Statistical Natural Language Processing, MIT Press.
  - Jurafsky & Martin (2000/2008):
    Speech and Language Processing, Prentice Hall.
  - Script (Lieblang) + Folien (Demberg)
- Skript und Folien auf der Kursseite oder in CLIX verfügbar

Kontakt

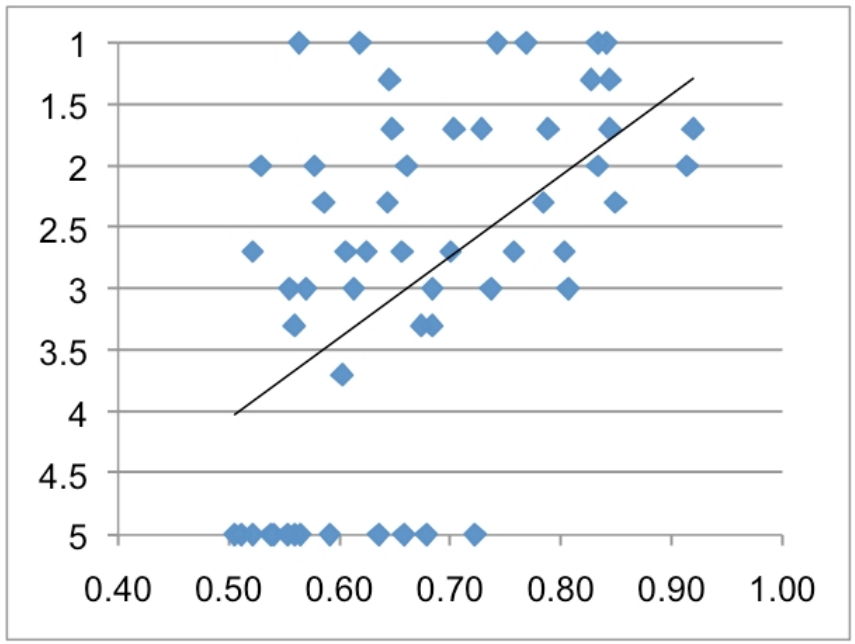
Bei Problemen bitte frühzeitig melden!

Kontaktadressen (bitte Termin vorher per E-Mail vereinbaren):

- Prof. Dr. Enrico Lieblang:
  `enrico.lieblang@htw.saarland.de`, Tel: 0681-5867 545
- Dr. Vera Demberg
  `vera@coli.uni-saarland.de`, Tel: 0681-302 70024

Prüfungsleistungen - Übungen

- Ca. 10–12 Übungsblätter
- Bedingungen für die einzelne Übungen variieren (u.a. Gruppenarbeit erlaubt oder nicht)
- 50% der Punkte von Übungen sind Voraussetzung für die Teilnahme an der Klausur



Prüfungsleistungen - Klausuren

- Zwischenklausur über den 1. Teil am 18. Juni 2012
- Endklausur: 26. Juli 2012
- Probeklausur die Woche davor
- Die Punkte von beiden Klausuren werden addiert, daraus gibt sich die Gesamtnote
- Anmeldung: siehe Information durch den Studienberater
- Ohne Anmeldung: Täuschversuch!

  Die Teilnahme an der Prüfung setzt die ordnungsgemässe Anmeldung zur Prüfung voraus. Die Teilnahme an der Prüfung bei versäumter Anmeldung wird als Täuschungsversuch gewertet und hat die Ungültigkeit der Prüfung zur Folge.

Accounts in der Computerlinguistik

"As a user of the CoLi computer systems it is mandatory that you check your email at regular intervals. Your CoLi email address is your official contact email address."

Account wird benötigt, um

- Information über Studium und Prüfungsmodalitäten vom Studienberater über den Verteiler zu bekommen
- Übungen einzureichen und den Kursleitern E-Mails zu schreiben
- Antrag:
  - Formular bei Frau Kröner ausfüllen (Prüfungssekretariat)
  - von Vera Demberg oder Stefan Thater unterschreiben lassen

Account-Einrichtung

- Siehe Wiki der Systemadministration
  `http://www.coli.uni-saarland.de/sg/`
- Passwortänderung:
  `http://wiki.coli.uni-saarland.de/wiki/index.php/Password`
- E-Mails weiterleiten:
  `http://wiki.coli.uni-saarland.de/wiki/index.php/Webmail-filters-forwards`
- Hilfe: Bei der Fachschaft anfragen
- Um auf manche Seiten zuzugreifen ist eine VPN-Verbindung nötig, siehe ITS-Seite:
  `http://www.its.uni-saarland.de/dienste/basisdienste/vpn/`
  (Uni-Kennung verwenden!)

Mailingliste

Mailingliste für Ankündigungen und Fragen von gemeinsamem Interesse:

`mathe3@ml.coli.uni-saarland.de`

Anmeldelink:
`http://ml.coli.uni-saarland.de/cgi-bin/mailman/listinfo/mathe3`

Abonnieren geht nur mit `...@coli.uni-saarland.de`!

CLIX—www.clix.uni-saarland.de

- E-Learning System der Universität
- Nicht mit HISPOS verwechseln (Klausuranmeldung)!
- Enthält die Materialien zum 1. Teil des Kurses
- Einloggen mit **Uni**-Kennung
- Kurs buchen (siehe nächste Folie)
- Skript befindet sich unter *Aktuelle Veranstaltungen* in *Lerninhalt*, Verteilungstabellen in der Bibliothek

Kurs in CLIX buchen

- In der linken Leiste auf *Vorlesungen* klicken
- Dann auf *Sommersemester 2012*, *Fakultät 4*, *4.7 Allgemeine Linguistik*, *Computerlinguistik*, *Kurse für B.Sc. Computerlinguistik*
- Rechts neben *Mathematische Grundlagen III - Statistische Methoden* auf den Einkaufskorb klicken
- Danach in der Leiste auf *Meine Kurse* und auf *Mein Warenkorb* gehen
- Ganz rechts unter *Aktion* auf das Symbol zum *Buchen* klicken
- Im aufgehenden Fenster *registrieren* wählen, dann Kurs starten

Ende des organisatorischen Teils

und worum soll's hier gehen?

Was ist statistische Sprachverarbeitung?

Was ist statistische Sprachverarbeitung?

- Anwendung von statistischen Methoden um Sprache zu verarbeiten
- maschinelles Lernen (überwacht, halbüberwacht, unüberwacht)
- Gegenteil: regelbasierte Sprachverarbeitung

Was brauchen wir an Statistik?

- Datenbeschreibung
- Zufallsvariablen
- schliessende Statistik

Korpora und datenintensive Linguistik

- Angenommen, wir wollen einen Parser bauen.
- **1. Möglichkeit**: wir schreiben Regeln, wie die Wörter zu Phrasen zusammengebaut werden.
  - Regeln von Hand zusammentragen: arbeitsintensiv
  - Interaktion zwischen Regeln
  - Ambiguität: "Ich sehe den Mann mit dem Fernrohr."
- **2. Möglichkeit**: wir lernen aus grossen Textmengen indem wir Regularitäten beobachten $\rightarrow$ Korpora
  - erlaubt Übergenerierung
  - alle Analysen werden bzgl. ihrer Wahrscheinlichkeit bewertet
  - beste Analyse finden
  - überwacht vs. unüberwacht: bessere Ergebnisse meist mit überwachten Methoden $\rightarrow$ *annotierte* Korpora

Typische Anwendungsgebiete statistischer Methoden

- Language Modelling
  - Spracherkennung
  - Rechtschreibkorrektur
  - POS Tagging: die richtige Wortart für jedes Wort bestimmen
- Parsing
  - Maschinelle Übersetzung
  - Disambiguierung
  - Informationsextraktion
- Klassifikation
  - Textkategorisierung
  - Spam Filter
  - Stimmungsanalyse (Sentiment Analysis)
- Clustering
  - Welche Texte / Wörter sind ähnlich oder verwandt

Methods

nützliche Methoden bei der Arbeit mit Korpora:

- Unix Werkzeuge
- Scripts
- Suchwerkzeuge

notwendige Statistik:

- Assoziationsmasse
- Statistische Tests
- Informationstheorie: Vorhersagbarkeit von Sprache

Geschichte der statistischen Sprachverarbeitung

- 1940er, frühe 1950er: sequentielle Modelle, Markovmodelle
- 1950er-1960er: Chomsky 1957 *"probabilistic models give no insight into the basic problems of syntactic structure"*
  1966 ALPAC report: keine Investition mehr in maschinelle Übersetzung
- 1970er-1980er: wenig statistische Arbeit im NLP Bereich, Ausnahme Fred Jelinek's Arbeitsgruppe bei IBM Watson; HMM und 3gram models (Spracherkennung, maschinelle Übersetzung)
- 1990er: Statistische Methoden werden zum dominanten Ansatz in der Computerlinguistik
- 2000er: drei verschiedene Communities (methodologisch)
  - traditionelle Ansätze
  - Anwendung einfacher statistische Methoden
  - Entwicklung von Methoden für maschinelles Lernen im Bereich der Sprachverarbeitung

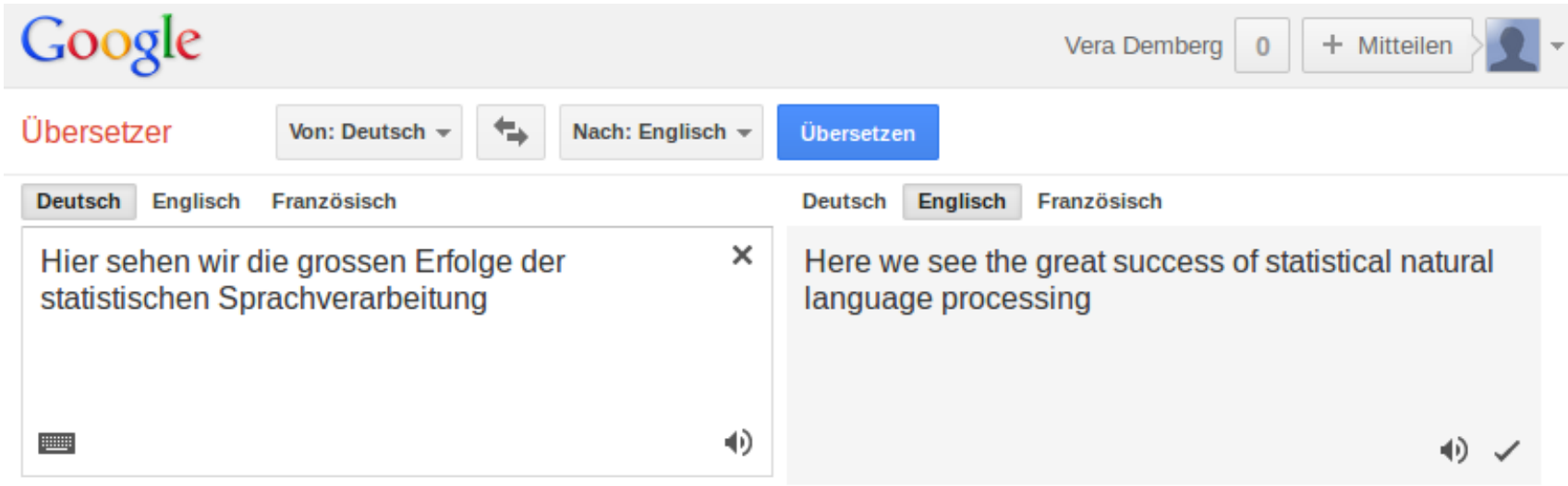
Erfolgsstories der statistischen Sprachverarbeitung

Siri. Beta

Your wish is its command.

Siri on iPhone 4S lets you use your voice to send messages, schedule meetings, place phone calls, and more. Ask Siri to do things just by talking the way you talk. Siri understands what you say, knows what you mean, and even talks back. Siri is so easy to use and does so much, you'll keep finding more and more ways to use it.

Erfolgsstories der statistischen Sprachverarbeitung



Erfolgsstories der statistischen Sprachverarbeitung

Ende

- nächster Termin: Montag, 16 Uhr. Thema: Motivation statistischer Ansätze
- Fragen?

(Folien dieser Vorlesung basieren auf Slides von Matt Crocker, Garance Paris, Hinrich Schütze)