Statistical Natural Language Processing

Prof. Dr. Dietrich Klakow

Lecture

- Lecture:
  - Friday 8:30-10:00
  - Location: MS Teams
  - Contact:
    - D. Klakow: dietrich.klakow@lsv.uni-saarland.de

Exercises

- Exercises:
  - First exercise sheet will be issued today
  - Three groups:
    - Vilém Zouhar
      - vzouhar@lsv.uni-saarland.de
      - Thursday 16:00-17:30
    - Awantee Deshpande
      - adeshpande@lsv.uni-saarland.de
      - Tuesday 14:15-15:45
    - Julius Steuer
      - jsteuer@coli.uni-saarland.de
      - Tuesday 12:15-13:45

Exercises

Submissions: Groups of 2 mandatory

Requirements:

- Threshold 70% (out of mandatory assignments)
- 10pts per assignment (around ~10) + bonus exercises (~2pts)
- 1pt bonus for active participation (showing up and **talking/contributing**)
- 5pts bonus for presenting the solution for an entire exercise sheet (at most once). Fractions for presenting parts of a sheet.

Exercises

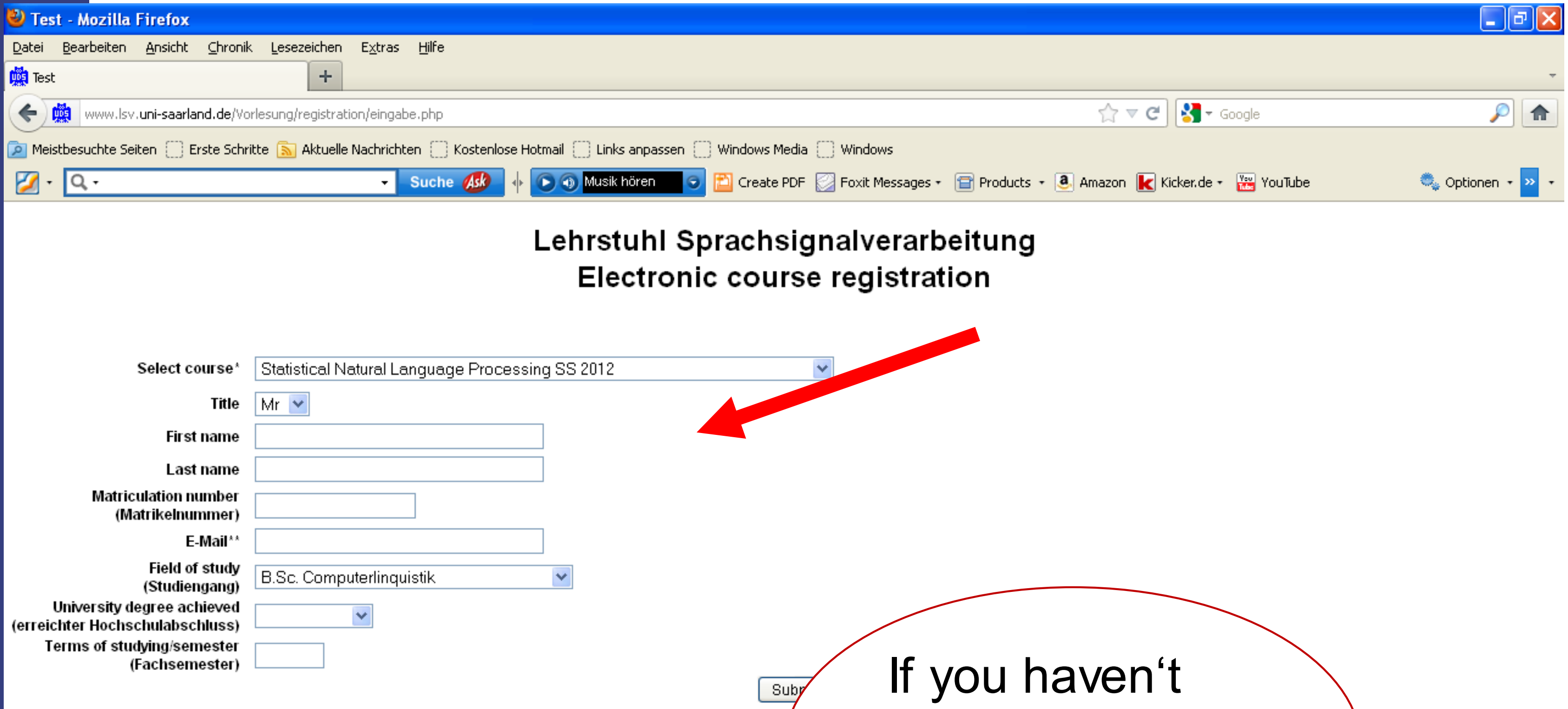
Submissions:

- Distribution of assignments using notebooks
- Suggestions: import your implementations rather than put it in the cells.
- Theoretical solutions also in the notebook (you can write LaTeX in cells)

Teams + Piazza

- Lectures and tutorials will be on MS Teams
- Forum: Piazza

Mailing List



Lehrstuhl Sprachsignalverarbeitung
Electronic course registration

Select course*	Statistical Natural Language Processing SS 2012
Title	Mr
First name	
Last name	
Matriculation number (Matrikelnummer)	
E-Mail**	
Field of study (Studiengang)	B.Sc. Computerlinquistik
University degree achieved (erreichter Hochschulabschluss)	
Terms of studying/semester (Fachsemester)	

If you haven‘t registered do so by today 11:00

Exam

- Friday, July 23rd, 8:00-10:00.
- GHH E22, Foyer GHH and HS002 in E13.
- Assignment to locations will follow in the week before the exam once July corona rules are clear

Literature

Foundations of Statistical Natural Language Processing

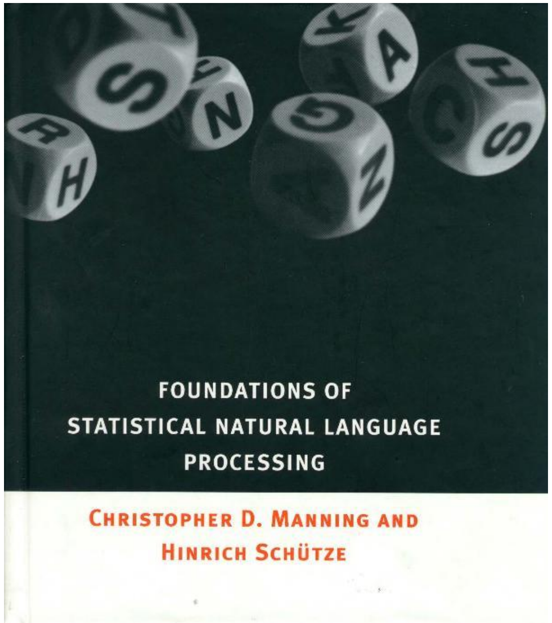
by Christopher D. Manning, Hinrich Schütze

Publisher: The MIT Press;

1st edition (June 18, 1999)

ISBN: 0262133601

List Price:$77.00

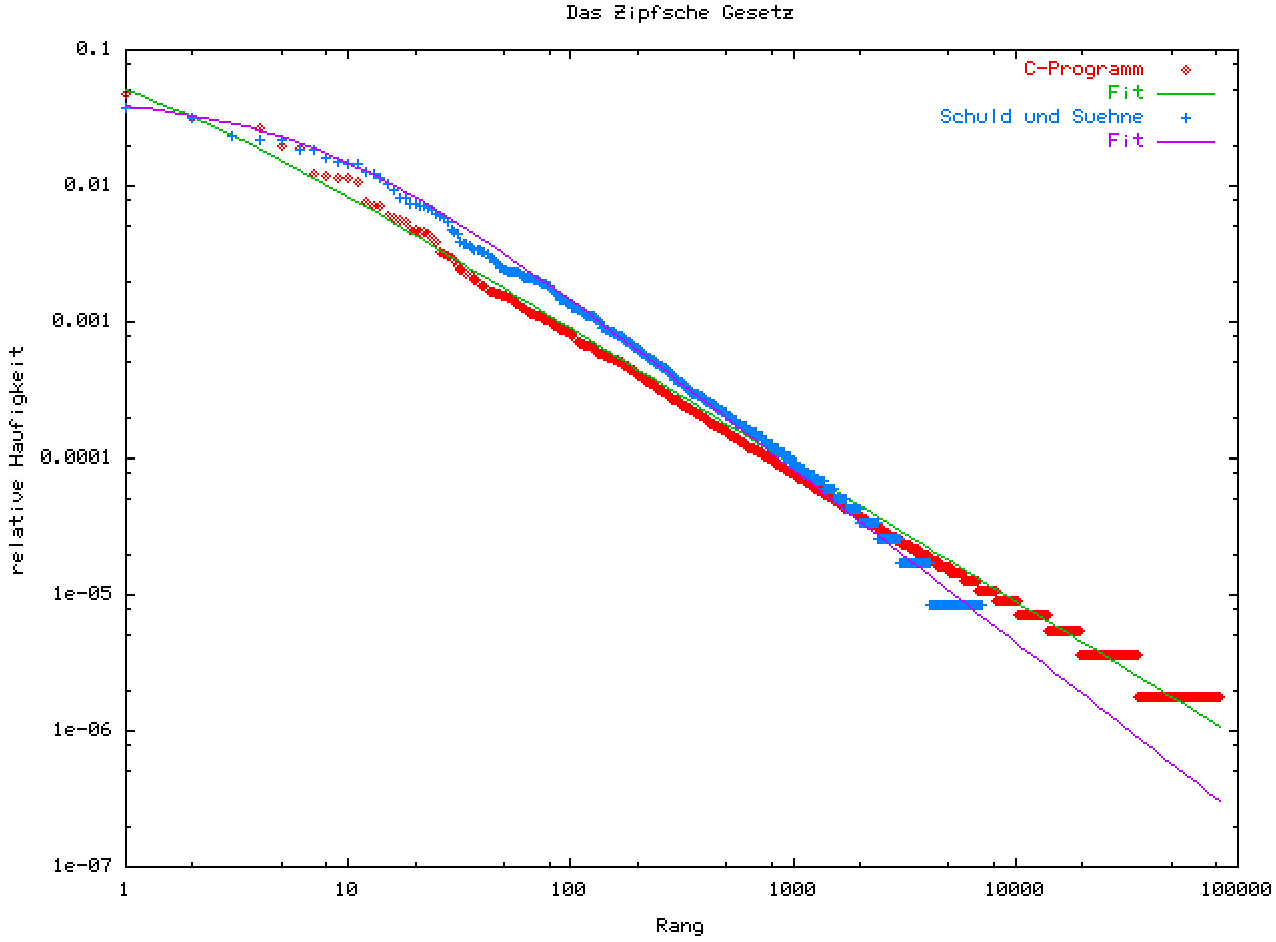


Rules of the Game

- In case you don´t understand something:
  1. Ask!!!
  2. Ask!!!
  3. Ask!!!

1. Introduction:
Overview of the Topics

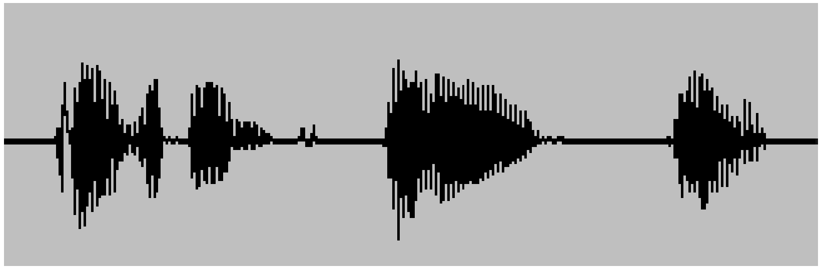
Use Zipf´s-Law in Language Modeling



Chapter 2: Natural Language as a Sequence of Symbols

- Zipf’s law
- Revision of basics of probability theory

Guess the next word

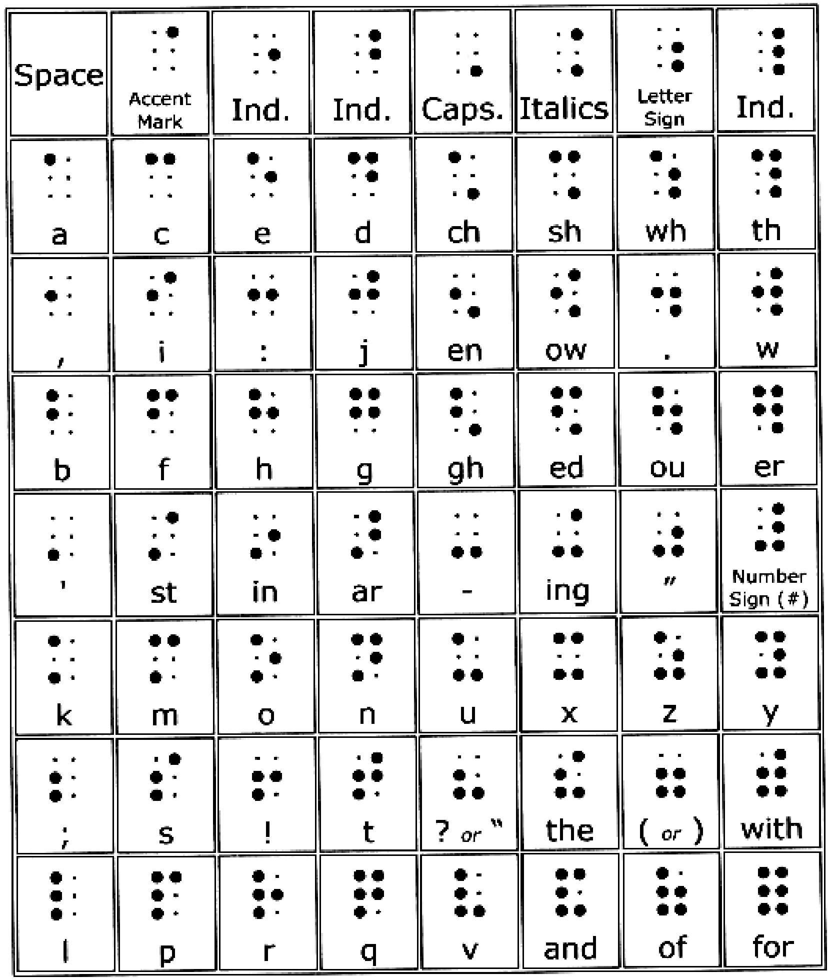


President Bill ???

Chapter 3: Basics of Language Modeling

- Language models for speech recognition
- Perplexity

Coding a language efficiently



⠀ Space	⠈ Accent Mark	⠐ Ind.	⠘ Ind.	⠠ Caps.	⠨ Italics	⠰ Letter Sign	⠸ Ind.
⠁ a	⠉ c	⠑ e	⠙ d	⠡ ch	⠩ sh	⠱ wh	⠹ th
⠂ ,	⠊ i	⠒ :	⠚ j	⠢ en	⠪ ow	⠲ .	⠺ w
⠃ b	⠋ f	⠓ h	⠛ g	⠣ gh	⠫ ed	⠳ ou	⠻ er
⠄ '	⠌ st	⠔ in	⠜ ar	⠤ -	⠬ ing	⠴ ”	⠼ Number Sign (#)
⠅ k	⠍ m	⠕ o	⠝ n	⠥ u	⠭ x	⠵ z	⠽ y
⠆ ;	⠎ s	⠖ !	⠞ t	⠦ ? *or* “	⠮ the	⠶ (*or*)	⠾ with
⠇ l	⠏ p	⠗ r	⠟ q	⠧ v	⠯ and	⠷ of	⠿ for

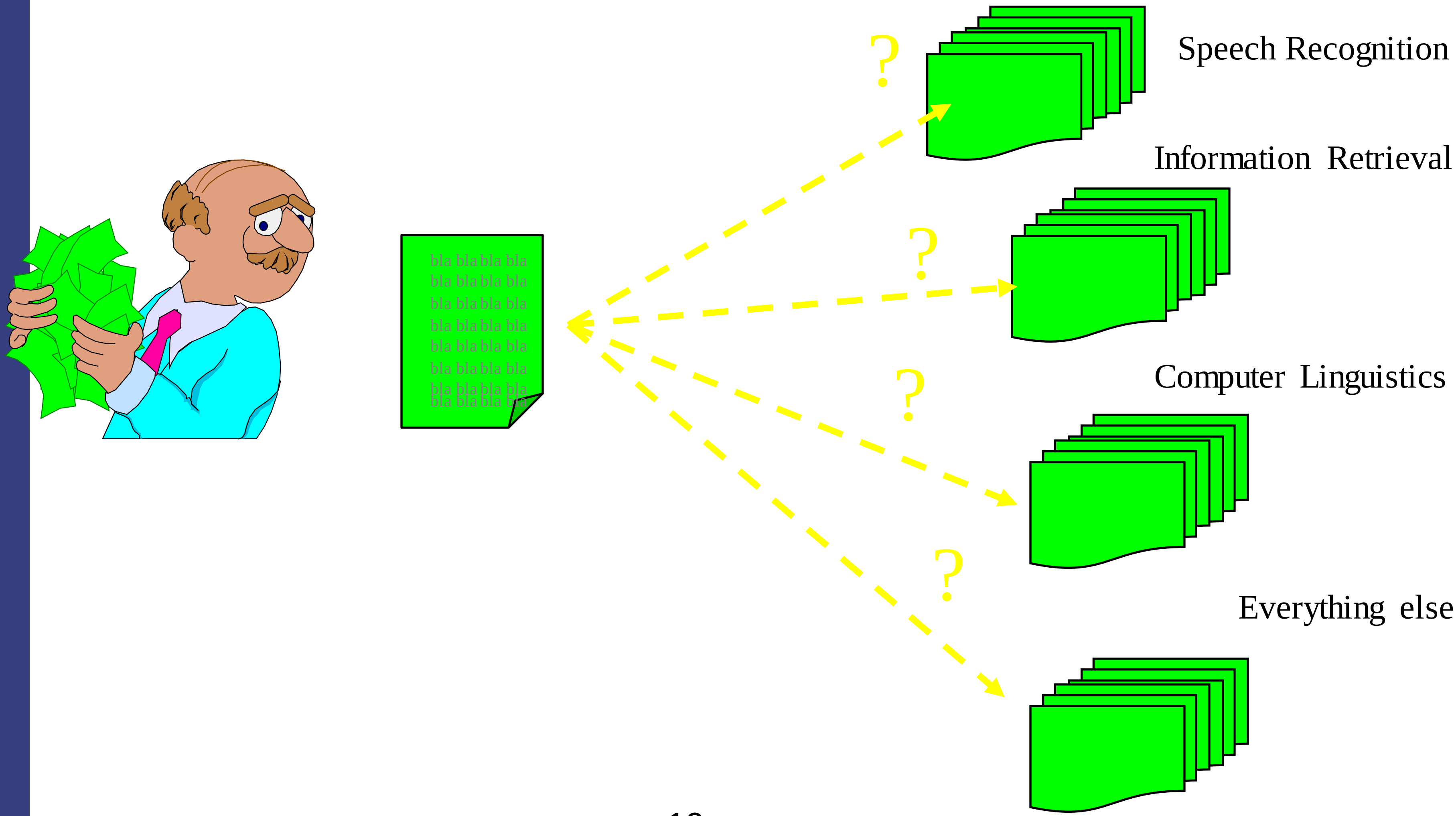
Chapter 4: Entropy

- The Shannon game
- Text compression

Chapter 5: Backing-Off Language Models

- Smoothing techniques

Text Categorization



Speech Recognition

Information Retrieval

Computer Linguistics

Everything else

Spam-Mail Classification

V/agra $3,3l

Amb/en

Mer/dia

C/alis $3,75

Val/um $l,21

X&nax

Som&

http://www.Chanatanxte.scriptmania.com/

Chapter 6. Text Classification

- Variants of Task
- Algorithms
    - Nearest Neighbor Classifier
    - Maximum Entropy Models
    - Decision Trees
    - Neural Networks
    - Unsupervised Clustering

Translation of the word „band“ into German (output from LEO’s)

- **band** das Band
- **band** die Band - *Musikgruppe*
- **band** [tech.] das Band
- **band** die Bandbreite
- **band** [chem.] die Bande - *im Spektrum* **band** das Beffchen **band** der Bereich **band** der Bund **band** der Frequenzbereich **band** die Gruppe **band** der Gurt **band** die Kapelle **band** die Leiste **band** die Musikkapelle **band** das Orchester **band** die Schar **band** die Schnur **band** [mus.] der Spielmannszug **band** der Streifen **band** die Truppe narrowband *also:* narrow-**band** *adj.* engbandig narrowband *also:* narrow-**band** *adj.* schmalbandig sideband *also:* side **band** [elec.] [telecom.] das Seitenband **Verben und Verbzusammensetzungen** to **band** together sich verbinden to **band** together sich vereinigen to **band** together sich zusammenrotten to **band** together sich zusammentun to **band** together zu einer Gruppe vereinigen to beat the **band** nie da gewesen sein to cross-**band** [tech.] absperren *[Holzverarbeitung]* **Zusammengesetzte Einträge** abrasive **band** - *cloth* [tech.] das Bandschleifleinen abrasive **band** - *paper* [tech.] das Bandschleifpapier adhesive **band** [tech.] das Klischeeklebeband attenuating **band** [aviat.] der Dämpfungsbereich audio **band** [phys.] der Hörbereich **band** aerial die Bandantenne **band**-aid das Heftpflaster **band**-aid [Amer.] [med.] das Pflaster **band**-aid [Amer.] [med.] das Wundpflaster **band** box die Hutschachtel **band** ceramics die Bandkeramik **band** collar der Stehkragen **band**-conveyor das Fließband **band** conveyor [tech.] der Gurtförderer **band**-conveyor das Transportband **band** edge die Bandkante **band** emission [autom.] die Bandemission **band** emission [autom.] die Bandenemission **band** gap [phys.] die Bandlücke **band** gate [tech.] der Bandausschnitt - *Spritzgusswerkzeug [Kunststoffe]* **band** grinder [tech.] die Bandschleifmaschine **band** matrix [math.] die Bandmatrix **band** of barrel das Fassband **band** of barrel der Fassreifen **band** of radiation [phys.] der Strahlungsbereich **band** of robbers die Räuberbande **band** overlap [tech.] die Bandüberlappung **band** printer [print.] der Banddrucker **band** radiation [autom.] die Bandenstrahlung **band** resaw [tech.] die Trennbandsäge **band** saw [tech.] die Bandsäge **band**-saw die Bandsäge **band** spectrum [tech.] das Bandenspektrum **band**-spread die Bandspreizung **band**-stand der Musikpavillon **band** structure [phys.] die Bandstruktur **band**-switch der Bereichsschalter **band**-switch der Bereichsumschalter **band** width die Bandbreite base **band** [tech.] das Basisband brake **band** [tech.] das Bremsband brass **band** [mus.] die Blaskapelle brass **band** [mus.] die Blechmusik brass **band** [mus.] der Spielmannszug broad **band** [tech.] das Breitband carrier **band** [tech.] das Trägerfrequenzband clay **band** [geol.] das Salband clincher **band** [autom.] das Wulstband *[Reifen]* conveyer **band** das Förderband cover **band** [tech.] das Deckband currency **band** [bank.] die Währungsbandbreite dance **band** die Tanzkapelle dead **band** [metr.] die Totzone edge **band** [tech.] der Umleimer *[Tischlerei]* elastic **band** [tech.] das Gummiband elastic **band** der Gummistrumpf error **band** der Zufallsstreubereich filter **band** [tech.] das Siebband flexible **band** die Randzeit - *Arbeitszeit* glassy **band** [tech.] glasiger Streifen guard **band** [elec.] der Rasen - *Abstand zwischen den Schrägspuren, den Videospuren, der benutzt wird, um eine gegenseitige Beeinflussung der Spuren zu vermeiden* guard **band** [elec.] der Schutzabstand - *Abstand zwischen den Schrägspuren, den Videospuren, der benutzt wird, um eine gegenseitige Beeinflussung der Spuren zu vermeiden* guard **band** [elec.] [telecom.] der Schutzbereich - *zwischen zwei Kanälen zur Vermeidung von Interferenzen* guard **band** [telecom.] der Schutzbereicht guard **band** [elec.] [telecom.] das Sicherheitsband - *zwischen zwei Kanälen zur Vermeidung von Interferenzen* guard **band** [elec.] [telecom.] das Sicherheitsfrequenzband - *zwischen zwei Kanälen zur Vermeidung von Interferenzen* guide **band** das Führungsband hair-**band** das Haarband heating **band** [tech.] das Heizband hinge **band** [tech.] das Gelenkband
  mehr >>

Chapter 7: Word Sense Disambiguation

- Dictionary-Based Disambiguation
- Thesaurus based methods
- Bayes Classifier

Example for Part-Of-Speech Tagging

Xinhua News Agency , Guangzhou , March 16 (Reporter Chen Ji) The latest statistics show that from January through February this year , the export of high-tech products in Guangdong Province reached 3.76 billion US dollars , up 34.8% over the same period last year and accounted for 25.5% of the total export in the province .

Example for Part-Of-Speech Tagging

Xinhua/NNP News/NNP Agency/NNP ,/, Guangzhou/NNP ,/, March/NNP 16/CD (/(Reporter/NNP Chen/NNP Ji/NNP)/SYM The/DT latest/JJS statistics/NNS show/VBP that/IN from/IN January/NNP through/IN February/NNP this/DT year/NN ,/, the/DT export/NN of/IN high-tech/JJ products/NNS in/IN Guangdong/NNP Province/NNP reached/VBD 3.76/CD billion/CD US/PRP dollars/NNS ,/, up/IN 34.8%/CD over/IN the/DT same/JJ period/NN last/JJ year/NN and/CC accounted/VBD for/IN 25.5%/CD of/IN the/DT total/JJ export/NN in/IN the/DT province/NN ./.

Chapter 8: Part-Of-Speech Tagging

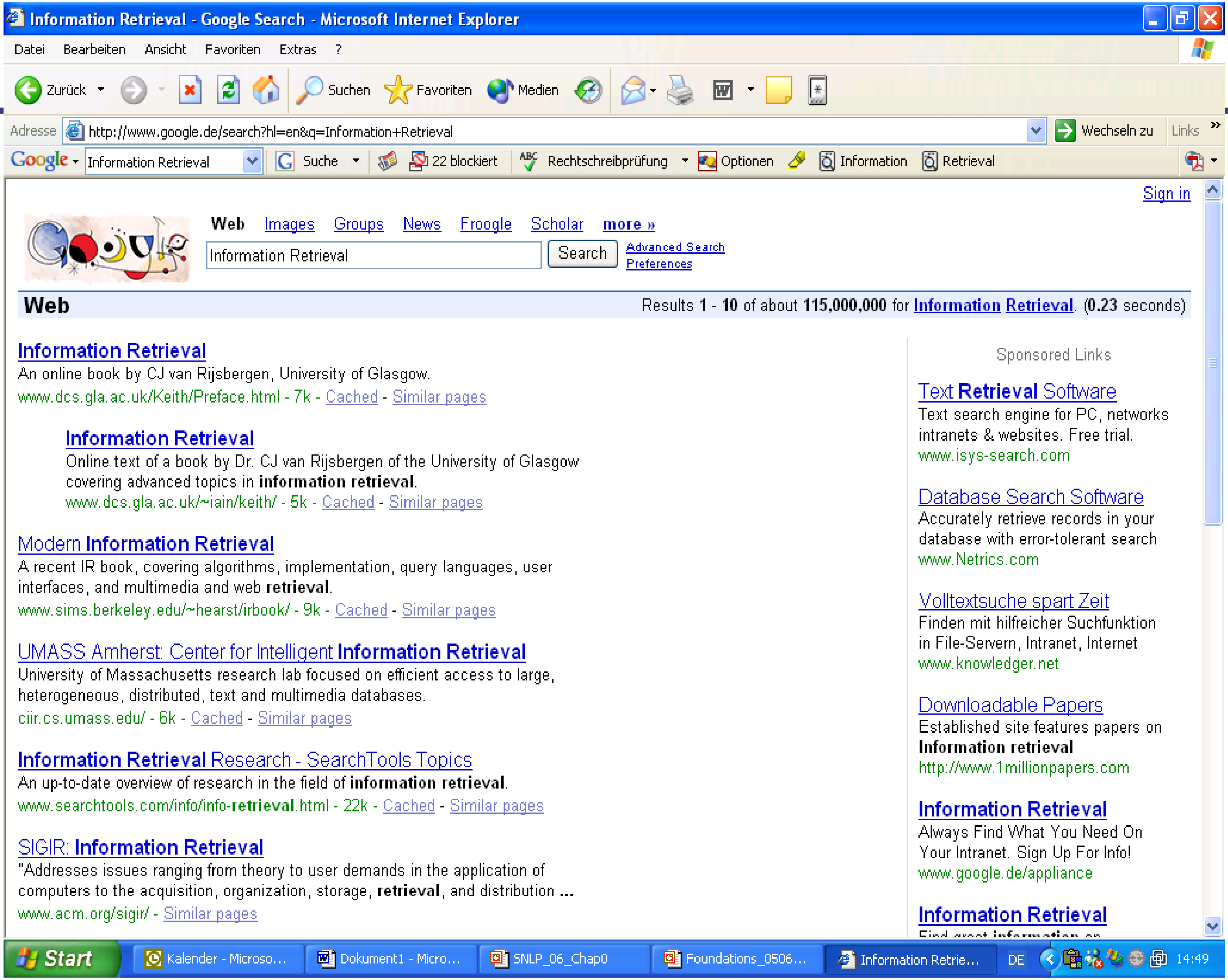
- Hidden Markov Model
- Rule based: The Brill tagger

Chapter 9. Named Entity Tagging

Task:

Identify names of people, organizations, locations … in text

- President <ENAMEX id="9" type="PERSON">Richard Nixon</ENAMEX> in <ENAMEX id="10" type="LOCATION">Moscow.</ENAMEX>



http://www.google.de/search?hl=en&q=Information+Retrieval

Sign in

Web Images Groups News Froogle Scholar more »

Information Retrieval | Search | Advanced Search | Preferences

Web Results **1** - **10** of about **115,000,000** for **Information Retrieval**. (**0.23** seconds)

Information Retrieval
An online book by CJ van Rijsbergen, University of Glasgow.
www.dcs.gla.ac.uk/Keith/Preface.html - 7k - Cached - Similar pages

Information Retrieval
Online text of a book by Dr. CJ van Rijsbergen of the University of Glasgow covering advanced topics in **information retrieval**.
www.dcs.gla.ac.uk/~iain/keith/ - 5k - Cached - Similar pages

Modern **Information Retrieval**
A recent IR book, covering algorithms, implementation, query languages, user interfaces, and multimedia and web **retrieval**.
www.sims.berkeley.edu/~hearst/irbook/ - 9k - Cached - Similar pages

UMASS Amherst: Center for Intelligent **Information Retrieval**
University of Massachusetts research lab focused on efficient access to large, heterogeneous, distributed, text and multimedia databases.
ciir.cs.umass.edu/ - 6k - Cached - Similar pages

Information Retrieval Research - SearchTools Topics
An up-to-date overview of research in the field of **information retrieval**.
www.searchtools.com/info/info-**retrieval**.html - 22k - Cached - Similar pages

SIGIR: **Information Retrieval**
"Addresses issues ranging from theory to user demands in the application of computers to the acquisition, organization, storage, **retrieval**, and distribution ...
www.acm.org/sigir/ - Similar pages

Sponsored Links

Text **Retrieval** Software
Text search engine for PC, networks intranets & websites. Free trial.
www.isys-search.com

Database Search Software
Accurately retrieve records in your database with error-tolerant search
www.Netrics.com

Volltextsuche spart Zeit
Finden mit hilfreicher Suchfunktion in File-Servern, Intranet, Internet
www.knowledger.net

Downloadable Papers
Established site features papers on **Information retrieval**
http://www.1millionpapers.com

Information Retrieval
Always Find What You Need On Your Intranet. Sign Up For Info!
www.google.de/appliance

Information Retrieval
Find great **information** on

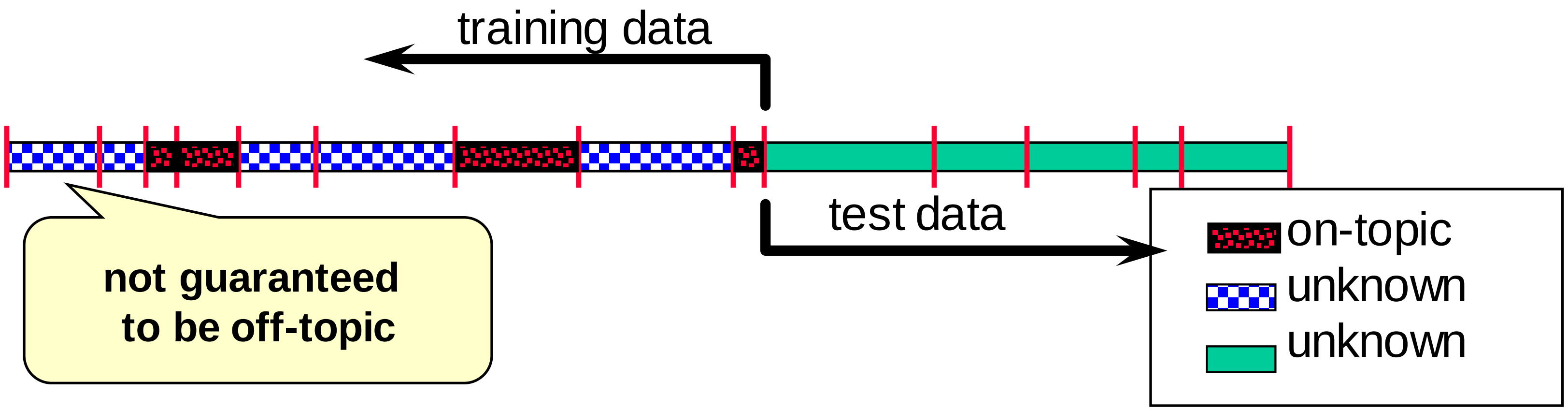
Chapter 10: Information Retrieval

- Evaluation
- Processing the query
- Vector space model
- Term weighting
- Distance metrics
- Models for term distribution
- Probabilistic IR
- Singular value decomposition
- Language models

Chapter 11. Topic Detection and Tracking (if time permits)

To detect stories that discuss the target topic, in multiple source streams.

- Find all the stories that discuss a given target topic
  - *Training:* Given N_t sample stories that discuss a given target topic,
  - *Test:* Find all subsequent stories that discuss the target topic.



training data

test data

not guaranteed to be off-topic

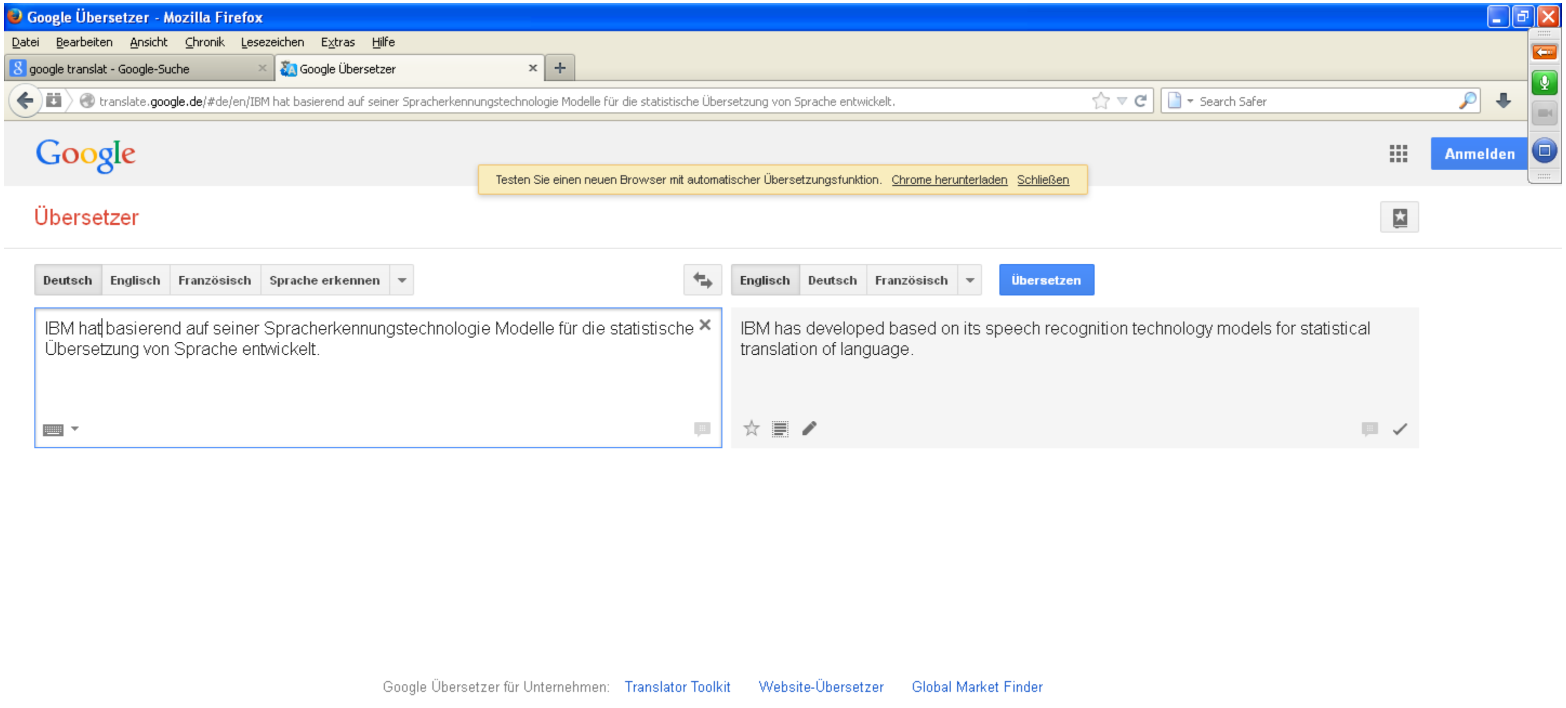
▨ on-topic

▦ unknown

🟩 unknown

Chapter 12: Statistical Machine Translation

- Machine translation as a sequence labeling problem
- IBM models 1-4



Google Übersetzer - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
google translat - Google-Suche
Google Übersetzer
Search Safer
Google
Anmelden
Übersetzer
Deutsch Englisch Französisch Sprache erkennen
Englisch Deutsch Französisch
Übersetzen

translate.google.de/#de/en/IBM hat basierend auf seiner Spracherkennungstechnologie Modelle für die statistische Übersetzung von Sprache entwickelt.

Testen Sie einen neuen Browser mit automatischer Übersetzungsfunktion. Chrome herunterladen Schließen

IBM hat basierend auf seiner Spracherkennungstechnologie Modelle für die statistische Übersetzung von Sprache entwickelt.

IBM has developed based on its speech recognition technology models for statistical translation of language.

Google Übersetzer für Unternehmen: Translator Toolkit Website-Übersetzer Global Market Finder

Summary Chapter 1

- Organization of the lecture
- Overview of the topics

- Are those the topics you are expecting?