Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Sprachtechnologie
Language Technology

Tobias Scheffer
Uwe Dick

Organization

- Lecture and exercise
- Lecture: Monday, 12:15-13:45, 03.04.0.02
- Exercise: Monday, 14:15-15:45, 3.04.0.02
- Homework is posted on our website each Tuesday and is due the following Monday.

Exam

- Successful completion of 70% of the homework exercises.
- 60 minutes of written exam, immediately followed by 15 mins of oral discussion.

Organization

- Website
  - Contains slides and in some cases online lecture videos.

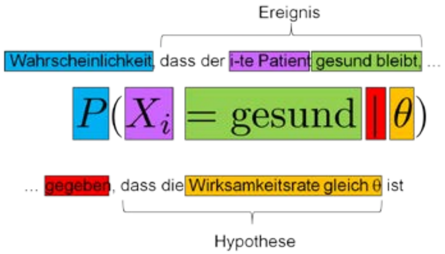
Literature

- Statistical natural language processing:
  - Manning & Schütze: „Foundations of Statistical Natural Language Processing.“ MIT Press
- Speech recognition
  - „The HTK Book“, im Internet verfügbar.
    - Speech Recognition Toolkit
    - http://htk.eng.cam.ac.uk/docs/docs.shtml
  - Huang, Acero und Hon: „Spoken Language Processing“. Prentice Hall.
- Information Retrieval:
  - Manning, Raghavan, Schütze: „Introduction to Information Retrieval“. Cambridge University Press.

Content

- Review of base technologies
  - Language models
  - Hiden Markov models, PCFG
  - Deep neural networks
- Language-processing tasks
  - Speech recognition
  - Translation, natural description of images
  - Parsing, information extraction
  - Text classification, clustering
- Information retrieval
  - Indexing and search

Mathematical Foundations



Ereignis

Wahrscheinlichkeit, dass der i-te Patient gesund bleibt, …

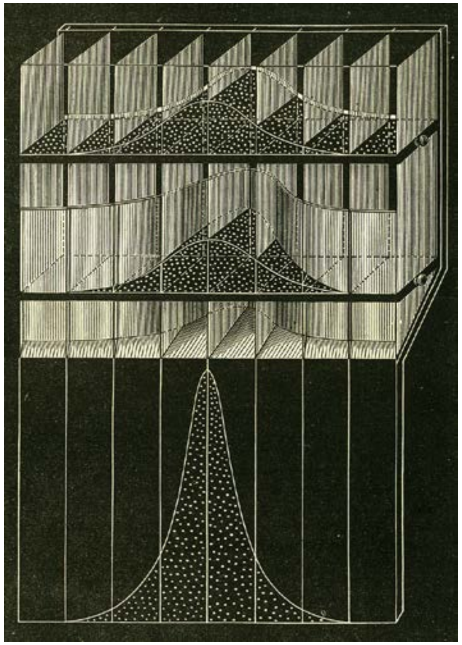
$$P(X_i = \text{gesund} \mid \theta)$$

… gegeben, dass die Wirksamkeitsrate gleich θ ist

Hypothese

$$P(Y|X) = P(X|Y)\frac{P(Y)}{P(X)}$$

$$\begin{aligned} P(X_{neu}|X_1,\ldots,X_n) &= \int_\theta P(X_{neu}|\theta, X_1,\ldots,X_n)P(\theta|X_1,\ldots,X_n)d\theta \\ &= \int_\theta P(X_{neu}|\theta)\ P(\theta|X_1,\ldots,X_n)\ d\theta \end{aligned}$$



Statistical Language Models

- Elementary tool for
  - Speech recognition
  - Spell checking
  - Auto-complete, machine translation, ….

- Quantifies probability of a sequence of terms in a language, with respect to a corpus

$$
\begin{aligned}
P(w_1,\ldots,w_T) &= P(w_1)P(w_2 \mid w_1)\ldots P(w_T \mid w_{T-1},\ldots,w_1) \\
&= P(w_1)P(w_2 \mid w_1)\ldots P(w_T \mid w_{T-1}, w_{T-N+1}) \\
&= \prod_{i=1}^{N-1} P(w_i \mid w_{i-1},\ldots,w_1) \prod_{i=N}^{T} P(w_i \mid w_{i-1},..w_{i-N+1})
\end{aligned}
$$

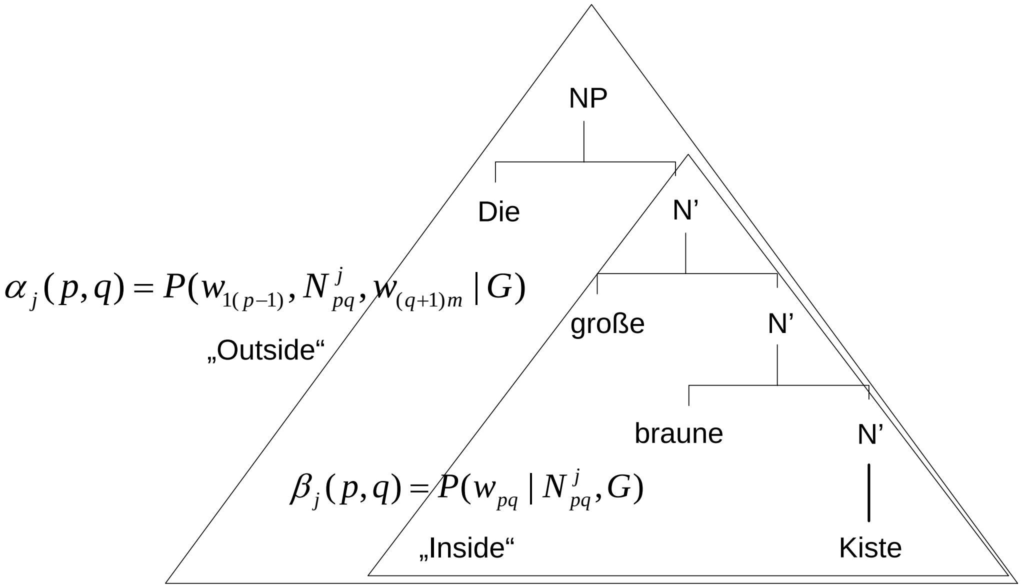
Scheffer/Dick: Sprachtechnologie

Statistical Language Models

- Grammar, acceptor, parser
  - Defines set of sentences of a language.
  - Defines hard boundaries.
  - Not a suitable mechanism for natural language.
  - Natural language has no hard boundaries, almost anything is possible.
- Statistical language model
  - Quantifies the probability of a sentence.
- Statistical parser
  - Infers the most likely syntactic interpretations of a sentence

Statistical Language Models

- N-Gram models
  - Sequence model with n-th order Markov assumption
- PCFG: probabilistic context-free grammar



$$\alpha_j(p,q) = P(w_{1(p-1)}, N^j_{pq}, w_{(q+1)m} \mid G)$$

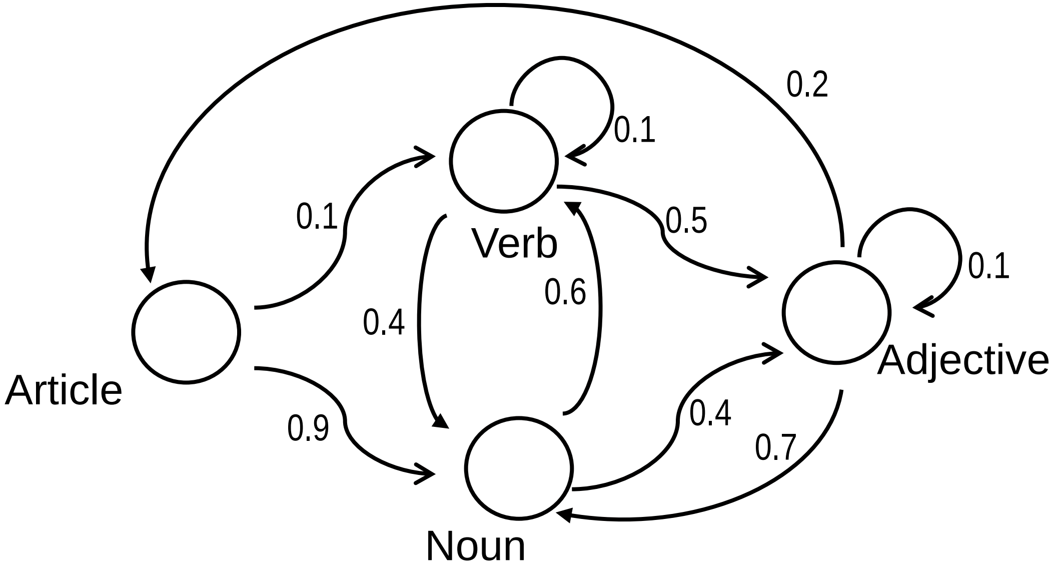
„Outside“

$$\beta_j(p,q) = P(w_{pq} \mid N^j_{pq}, G)$$

„Inside“

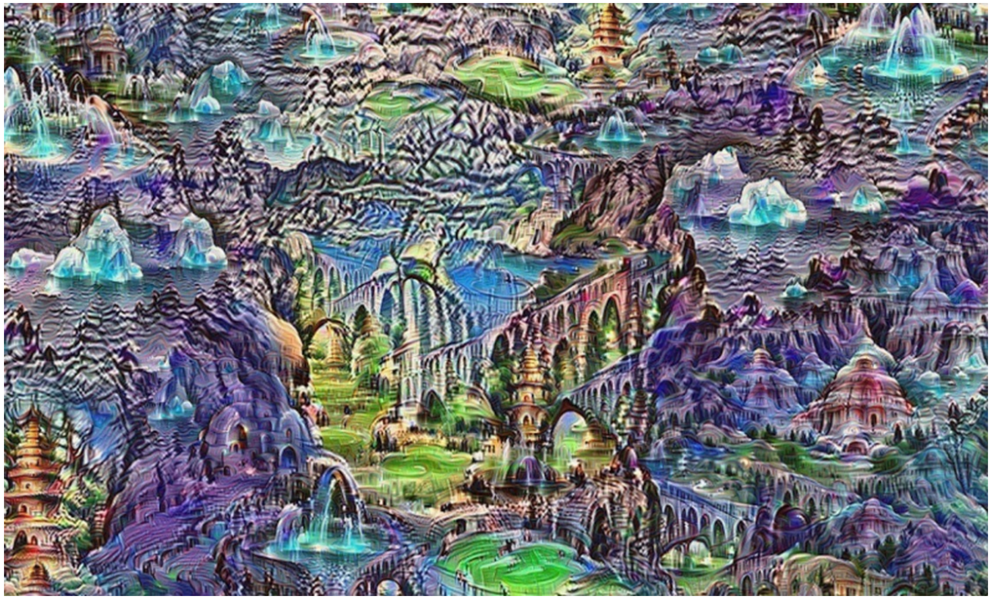
Hidden-Markov-Modell

- Probabilistic model for sequences.
- Used, for instance, in
  - speech recognition,
  - part-of-speech tagging.



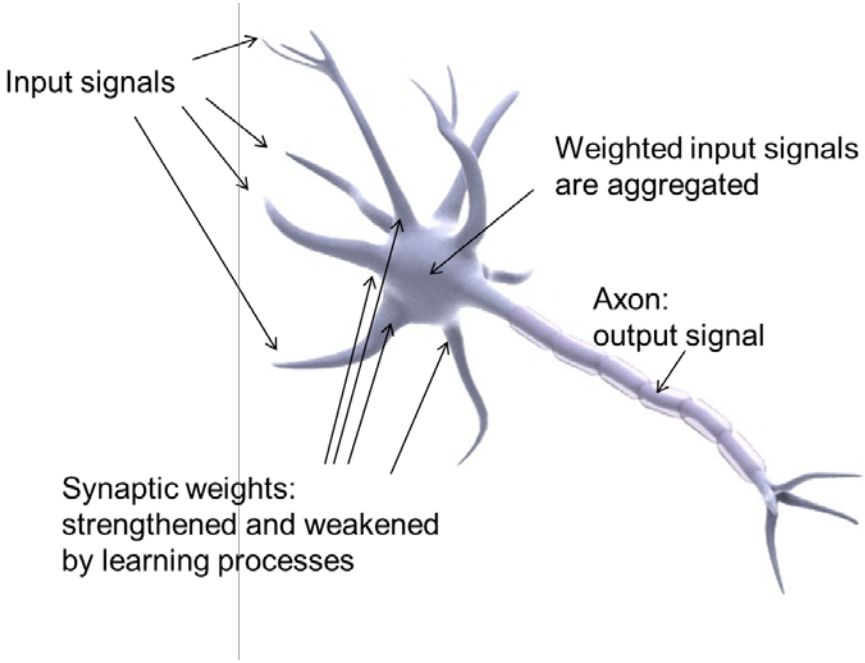
Deep Neural Networks

- Computational models of neural information processing.
- Good at learning abstract feature representations of complex input objects



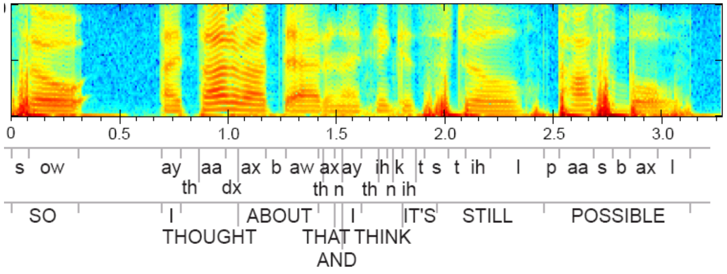
Deep Neural Networks

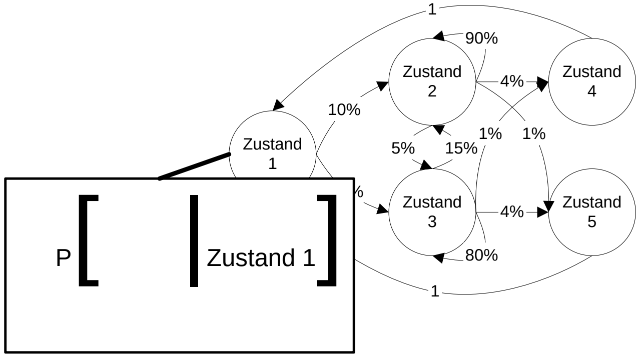
- State of the art for
  - Speech recognition
  - Translation
  - Natural description of images
  - …


Input signals
Weighted input signals are aggregated
Axon: output signal
Synaptic weights: strengthened and weakened by learning processes

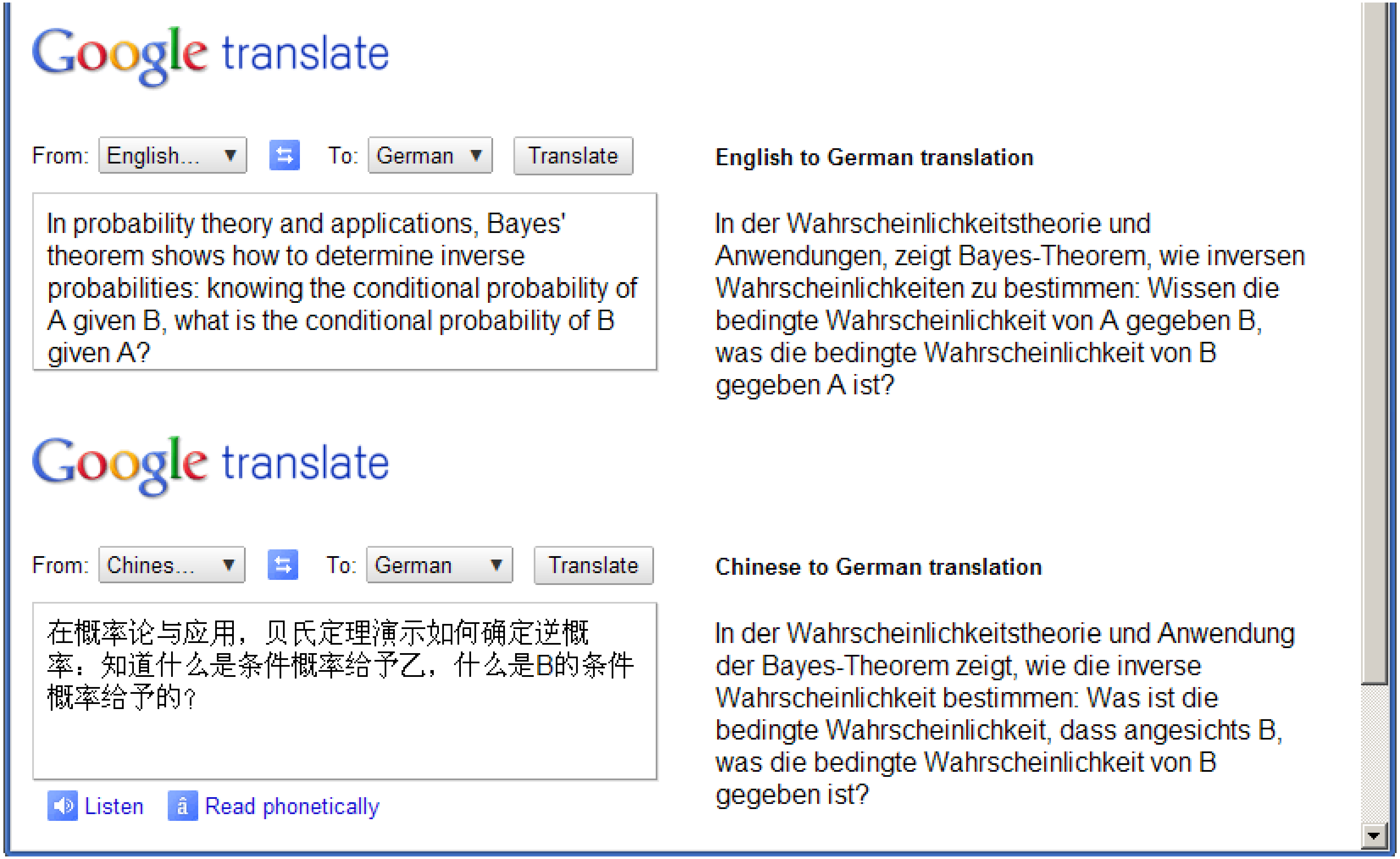
Speech Recognition

- Statistical speech recognition
  - Acoustic model +
  - Language model.
- Neural networks for speech recognition
- Text to speech





Machine Translation



From: English... To: German

In probability theory and applications, Bayes' theorem shows how to determine inverse probabilities: knowing the conditional probability of A given B, what is the conditional probability of B given A?

English to German translation

In der Wahrscheinlichkeitstheorie und Anwendungen, zeigt Bayes-Theorem, wie inversen Wahrscheinlichkeiten zu bestimmen: Wissen die bedingte Wahrscheinlichkeit von A gegeben B, was die bedingte Wahrscheinlichkeit von B gegeben A ist?

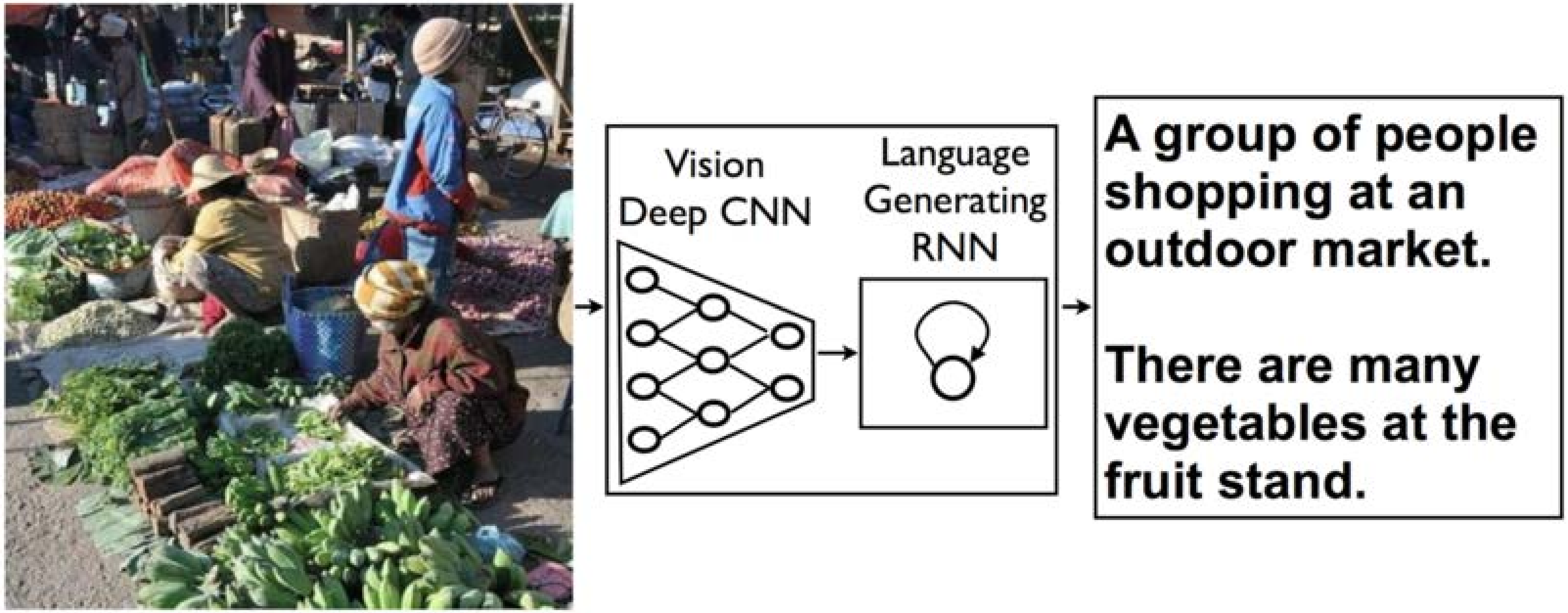
From: Chines... To: German

在概率论与应用，贝氏定理演示如何确定逆概率：知道什么是条件概率给予乙，什么是B的条件概率给予的?

Chinese to German translation

In der Wahrscheinlichkeitstheorie und Anwendung der Bayes-Theorem zeigt, wie die inverse Wahrscheinlichkeit bestimmen: Was ist die bedingte Wahrscheinlichkeit, dass angesichts B, was die bedingte Wahrscheinlichkeit von B gegeben ist?

Text Generation, Description of Images

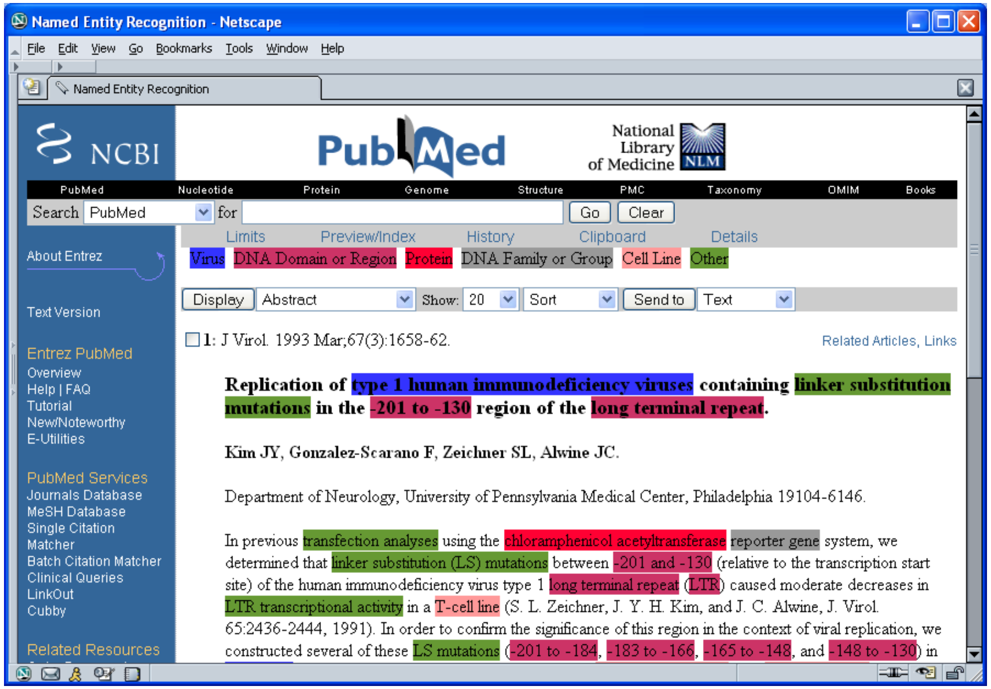


A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

POS-Tagging, Named Entity Recognition

- Identifying parts of speech
- Identifying proper names of persons, organizations, places, times, dates, genes, molecules, …


Named Entity Recognition - Netscape
File Edit View Go Bookmarks Tools Window Help
Named Entity Recognition
NCBI
PubMed
National Library of Medicine NLM
PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books
Search PubMed for Go Clear
Limits Preview/Index History Clipboard Details
Virus DNA Domain or Region Protein DNA Family or Group Cell Line Other
Display Abstract Show: 20 Sort Send to Text
About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby
Related Resources
Related Articles, Links

□ 1: J Virol. 1993 Mar;67(3):1658-62.

Replication of type 1 human immunodeficiency viruses containing linker substitution mutations in the -201 to -130 region of the long terminal repeat.

Kim JY, Gonzalez-Scarano F, Zeichner SL, Alwine JC.

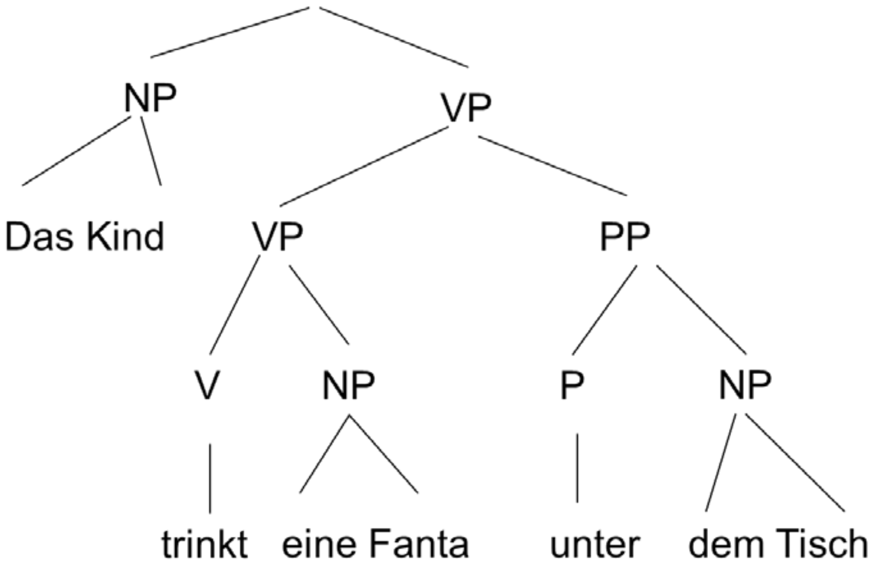
Department of Neurology, University of Pennsylvania Medical Center, Philadelphia 19104-6146.

In previous transfection analyses using the chloramphenicol acetyltransferase reporter gene system, we determined that linker substitution (LS) mutations between -201 and -130 (relative to the transcription start site) of the human immunodeficiency virus type 1 long terminal repeat (LTR) caused moderate decreases in LTR transcriptional activity in a T-cell line (S. L. Zeichner, J. Y. H. Kim, and J. C. Alwine, J. Virol. 65:2436-2444, 1991). In order to confirm the significance of this region in the context of viral replication, we constructed several of these LS mutations (-201 to -184, -183 to -166, -165 to -148, and -148 to -130) in

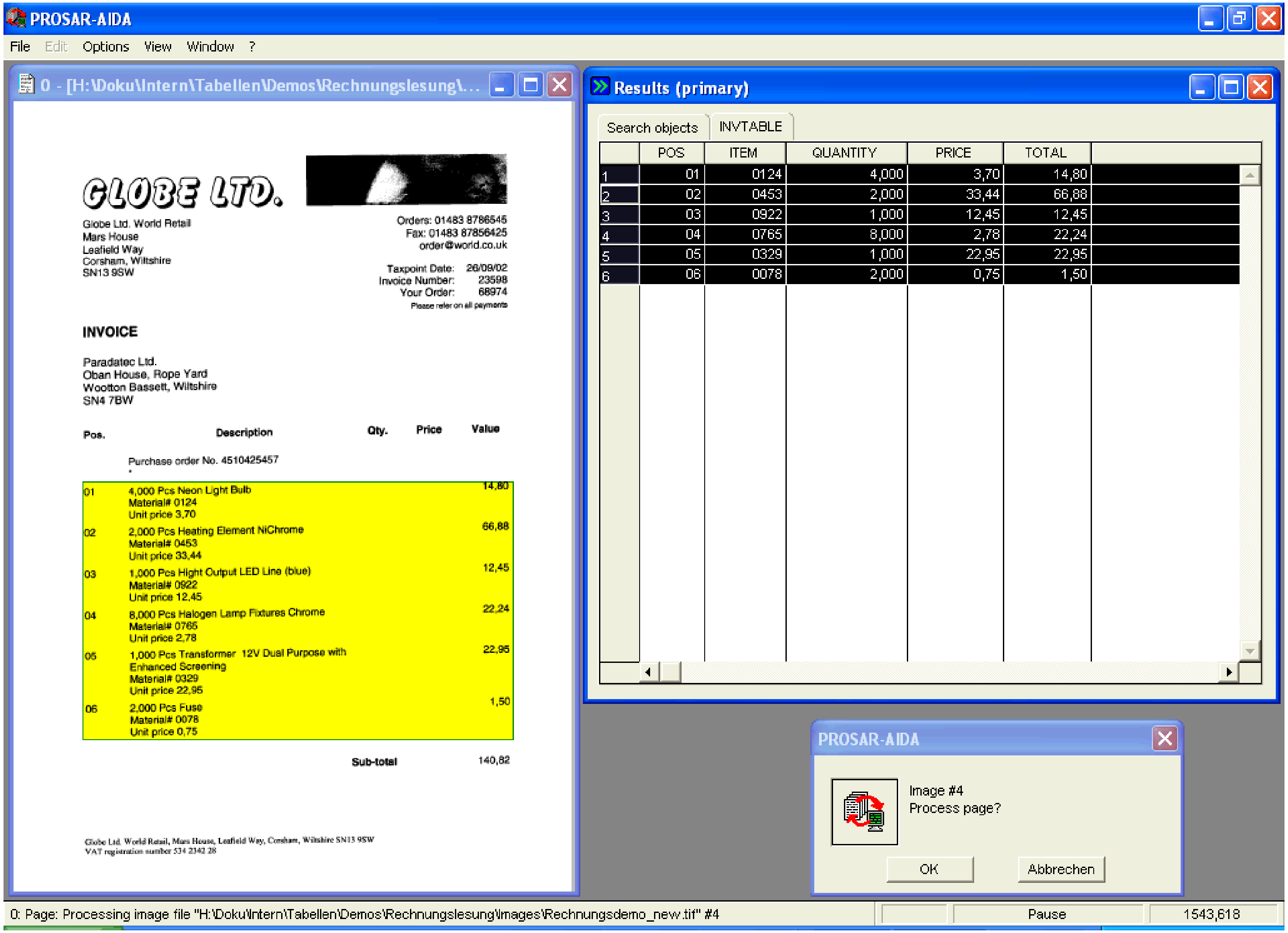
Scheffer/Dick: Sprachtechnologie

Parsing

- Finding the most likely syntactical structure of sentences.



Text Classification, Information Extraktion


PROSAR-AIDA
File Edit Options View Window ?
0 - [H:\Doku\Intern\Tabellen\Demos\Rechnungslesung\...
Results (primary)
Search objects
INVTABLE

GLOBE LTD.

Globe Ltd. World Retail
Mars House
Leafield Way
Corsham, Wiltshire
SN13 9SW

Orders: 01483 8786545
Fax: 01483 87856425
order@world.co.uk

Taxpoint Date: 26/09/02
Invoice Number: 23598
Your Order: 68974
Please refer on all payments

INVOICE

Paradatec Ltd.
Oban House, Rope Yard
Wootton Bassett, Wiltshire
SN4 7BW

Pos.	Description	Qty.	Price	Value
	Purchase order No. 4510425457 *			
01	4,000 Pcs Neon Light Bulb Material# 0124 Unit price 3,70			14,80
02	2,000 Pcs Heating Element NiChrome Material# 0453 Unit price 33,44			66,88
03	1,000 Pcs Hight Output LED Line (blue) Material# 0922 Unit price 12,45			12,45
04	8,000 Pcs Halogen Lamp Fixtures Chrome Material# 0765 Unit price 2,78			22,24
05	1,000 Pcs Transformer 12V Dual Purpose with Enhanced Screening Material# 0329 Unit price 22,95			22,95
06	2,000 Pcs Fuse Material# 0078 Unit price 0,75			1,50
		Sub-total		140,82

Globe Ltd. World Retail, Mars House, Leafield Way, Corsham, Wiltshire SN13 9SW
VAT registration number 534 2342 28

	POS	ITEM	QUANTITY	PRICE	TOTAL
1	01	0124	4,000	3,70	14,80
2	02	0453	2,000	33,44	66,88
3	03	0922	1,000	12,45	12,45
4	04	0765	8,000	2,78	22,24
5	05	0329	1,000	22,95	22,95
6	06	0078	2,000	0,75	1,50

PROSAR-AIDA

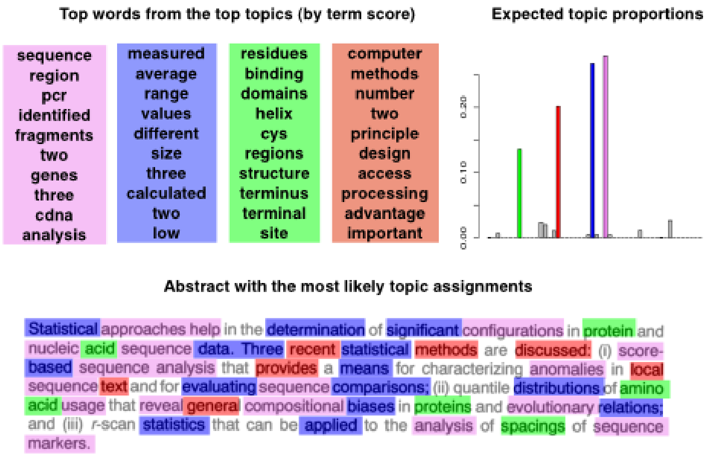
Image #4
Process page?

OK Abbrechen

0: Page: Processing image file "H:\Doku\Intern\Tabellen\Demos\Rechnungslesung\Images\Rechnungsdemo_new.tif" #4 | Pause | 1543,618

Clustering and Topic Models

- Clustering: grouping related documents.
- Topic modeling: groupung related terms in documents.



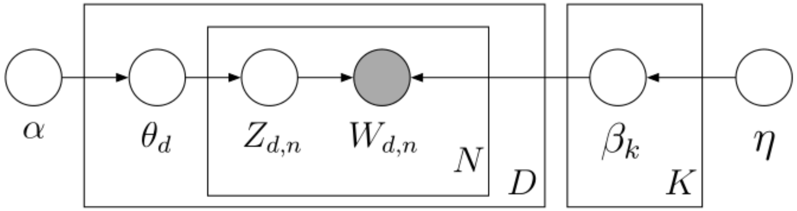
Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions

Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.



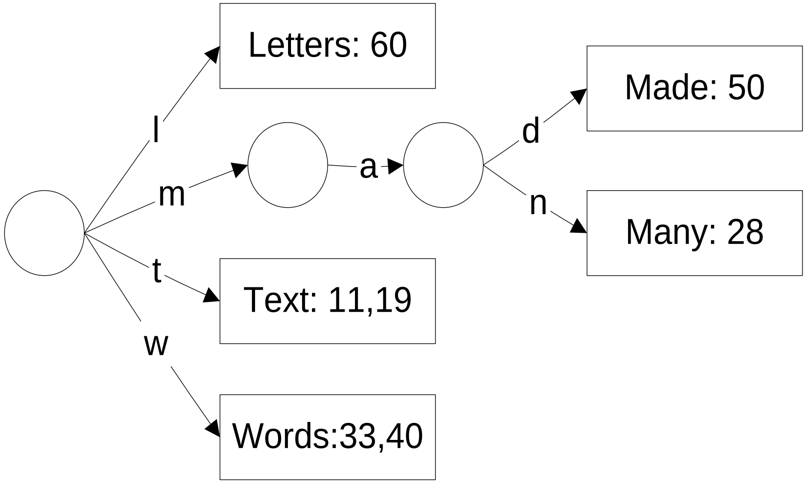
Indexing and Search

- Fast search in large text corpora.

1 6 9 11 17 19 24 28 33 40 46 50 55 60

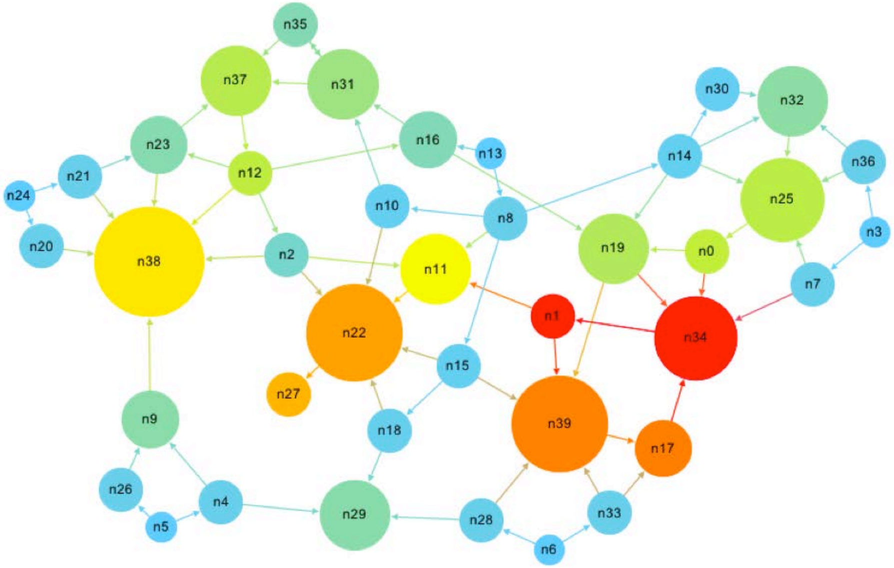
This is a text. A text has many words. Words are made from letters.

Terme	Vorkommen
Letters	60
Made	50
Many	28
Text	11, 19
words	33, 40



Web Search

- Crawling: visit which URL when? Challenges:
  - Infinite URLs, dynamic page content.
  - Update frequency and schedules.
  - Identical pages with multiple URLs.
  - Link spam.
- Relevance ranking,
  link analysis.



Questions?

- No exercise sheet this week.
- No exercise this week.