Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Sprachtechnologie

Tobias Scheffer, Tom Vanck, Paul Prasse

Organisation

- Vorlesung/Übung, praktische Informatik.
- 4 SWS.
- Termin:
  - Montags, 10-11:30, 03.04.0.02 (ab 03.05.)
  - Montags, 12-13:30, 03.04.0.02

- Heute S21

Orgnisation

- Diplom, Bachelor, Master.
- Ab 5. Semester empfohlen.

Organisation

- Webseite.
- Kalender.
  - Vorlesungs- und Übungstermine.
- Blog:
  - Ihre Fragen, Kommentare.
- Folien:
  - Am Tag nach der Vorlesung im Netz.

Organisation

- Übungsaufgaben:
  - Am Tag nach der Vorlesung im Netz.
  - Werden in der darauffolgenden Übung besprochen.
  - Sie können für einzelne Aufgaben votieren.
  - Sie müssen für 2/3 der Aufgaben des Semesters votieren, um die Prüfung abzulegen.
  - Sie rechnen votierte Aufgaben vor.
- Mündliche Prüfung am Ende des Semesters.

Literatur

- Folienkopien auf der Webseite
- Statistische Sprachverarbeitung:
  - Manning & Schütze: „Foundations of Statistical Natural language Processing.“ MIT Press
- Spracherkennung:
  - „The HTK Book“, im Internet verfügbar.
  - Huang, Acero und Hon: Spoken Language Processing. Prentice Hall.
- Information Retrieval:
  - Christopher Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval. Cambridge University Press.

Inhalt

- Verarbeitung geschriebener und gesprochener natürlicher Sprache.
  - Spracherkennung, Sprachportale,
  - Klassifikation, Informationsextraktion.
  - Information Retrieval, Suche, Websuche.

Mathematische Grundlagen

Zufallsvariablen

- Ein Experiment ist ein definierter Prozess, in dem eine Beobachtung erzeugt wird.
- Ereignisraum Ω: Alle möglichen Ausgänge
- Zufallsvariable X: Abbildung des Ereignisraumes auf numerische Werte. $P(X=x) = P(A \mid X(A)=x)$.
- Wahrscheinlichkeitsfunktion P verteilt Wahrscheinlichkeitsmasse 1 auf Elemente in Ω.
- Sicheres Ereignis: $P(X \in \Omega_X)=1$.
- Unmögliches Ereignis: $P(X \in \varnothing)=0$.
- Mathematische Grundlage durch Kolmogoroff-Axiome.

Multivariate Normalverteilung

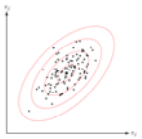
- Merkmalsvektoren x und Mittelwertvektor μ haben d Dimensionen.
- Korvarianzmatrix Σ (Größe d x d).

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

- Bedeutung von Mittelwert und Kovarianz:

$$\mu_i = \mathcal{E}[x_i] \qquad \sigma_{ij} = \mathcal{E}[(x_i-\mu_i)(x_j-\mu_j)]$$

- Wie sieht die Kovarianzmatrix aus?



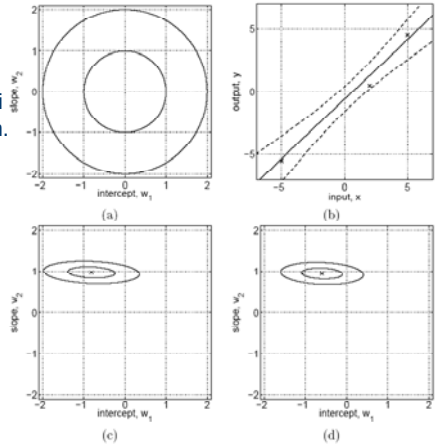
Log-Likelihood

- Wie wahrscheinlich sind die Daten gegeben das Modell?
  - $-\log P(L \mid f_{\mathbf{w}}) = -\log P(y_1,...,y_N \mid f_{\mathbf{w}}, \mathbf{x}_1,...,\mathbf{x}_N)$
- Annahme: Datenpunkte sind unabhängig gezogen.
  - $-\log P(y_1,...,y_N \mid f_{\mathbf{w}}, \mathbf{x}_1,...,\mathbf{x}_N)$
    $= -\log \prod_i P(y_i \mid f_{\mathbf{w}}, \mathbf{x}_i)$
    $= \sum_i -\log P(y_i \mid f_{\mathbf{w}}, \mathbf{x}_i)$
    $= \frac{1}{\lambda'} \sum_i l(f_{\mathbf{w}}(\mathbf{x}_i), y_i)$

Annahme: spezielle Exponential-verteilung

Bayessche Regression

- (a) Prior $P(\mathbf{w})$
- (b) Regressionsgerade, $\overline{\mathbf{w}}$; Korridor von zwei Standardabweichungen.
- (c) Likelihood $P(\mathbf{y} \mid X, \mathbf{w})$
- (d) Posterior $P(\mathbf{w} \mid X, \mathbf{y})$



Statistische Sprachmodelle

- Elementares Werkzeug für
  - Spracherkennung,
  - Rechntschreibkorrektur,
  - Auto-Complete, Übersetzung, …

- Wahrscheinlichkeit einer Abfolge von Wörtern.
  - „Ich pflücke Beeren“ vs. „Ich pflücke Bären“.

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

Statistische Sprachmodelle

- Grammatik, Akzeptor, Parser:
  - Menge der Sätze einer Sprache.
  - Als Mechanismus für Verarbeitung natürlicher Sprache nicht geeignet.
  - Sprache hat keine scharfen Ränder, fast alles ist möglich.
- Statistisches Sprachmodell, statistische Inferenz.
  - Wahrscheinlichkeit eines Satzes.
  - Wahrscheinlichste Interpretation.

Markov-Prozesse

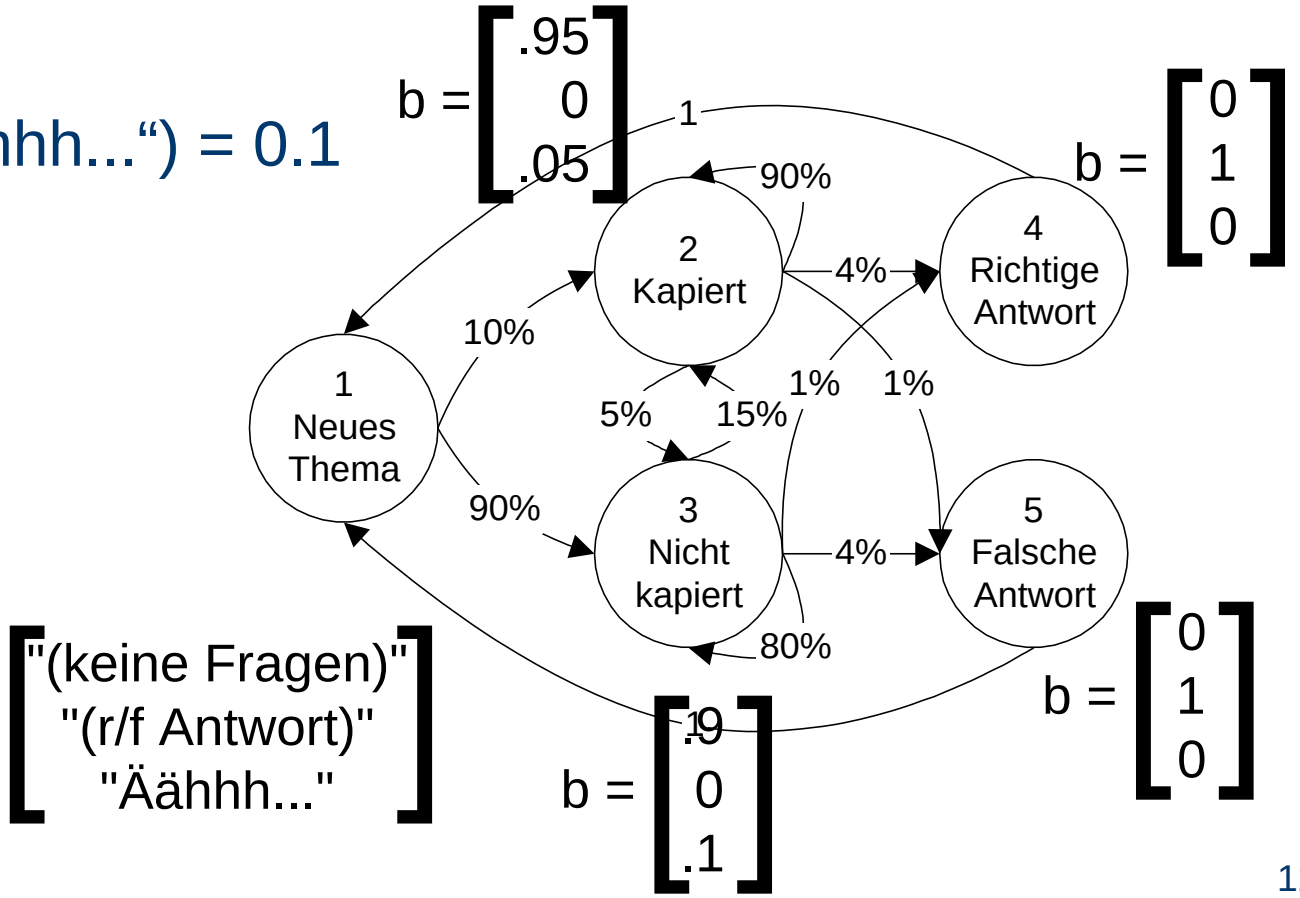
- X_1, ..., X_n: Zufallsvariablen.
- Allgemein gilt: $P(X_1,...,X_n) = P(X_1)\prod_{i=2}^{n} P(X_i \mid X_{i-1},...,X_1)$
- Zufallsvariablen bilden eine Markovkette, gdw:
  $P(X_1,...,X_n) = P(X_1)\prod_{i=2}^{n} P(X_i \mid X_{i-1})$
- Jede Variable X_i nur von Vörgänger X_{i-1} abhängig.

- Morkov-Modell:
  Probabilistischer endlicher Automat, Folge der Zustände ist Markov-Kette.

- (Andrei Markov, 1856-1922)

Hidden-Markov-Modell

- Akustisches Modell für Spracherkennung.
- Zustände emittieren Beobachtungen O_t (mit Wahrscheinlichkeit $b_i(O_t)$).
- b_3(„Äähhh...“) = 0.1

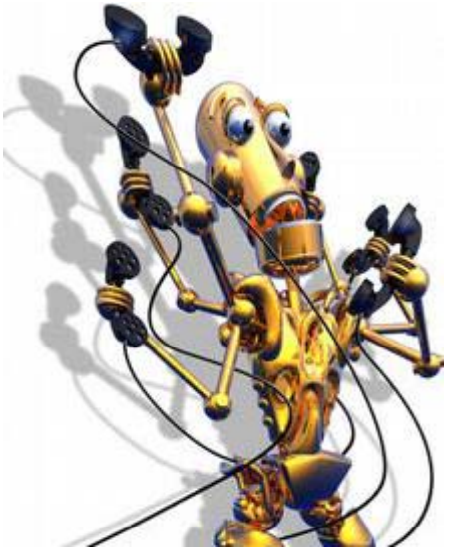


Spracherkennung

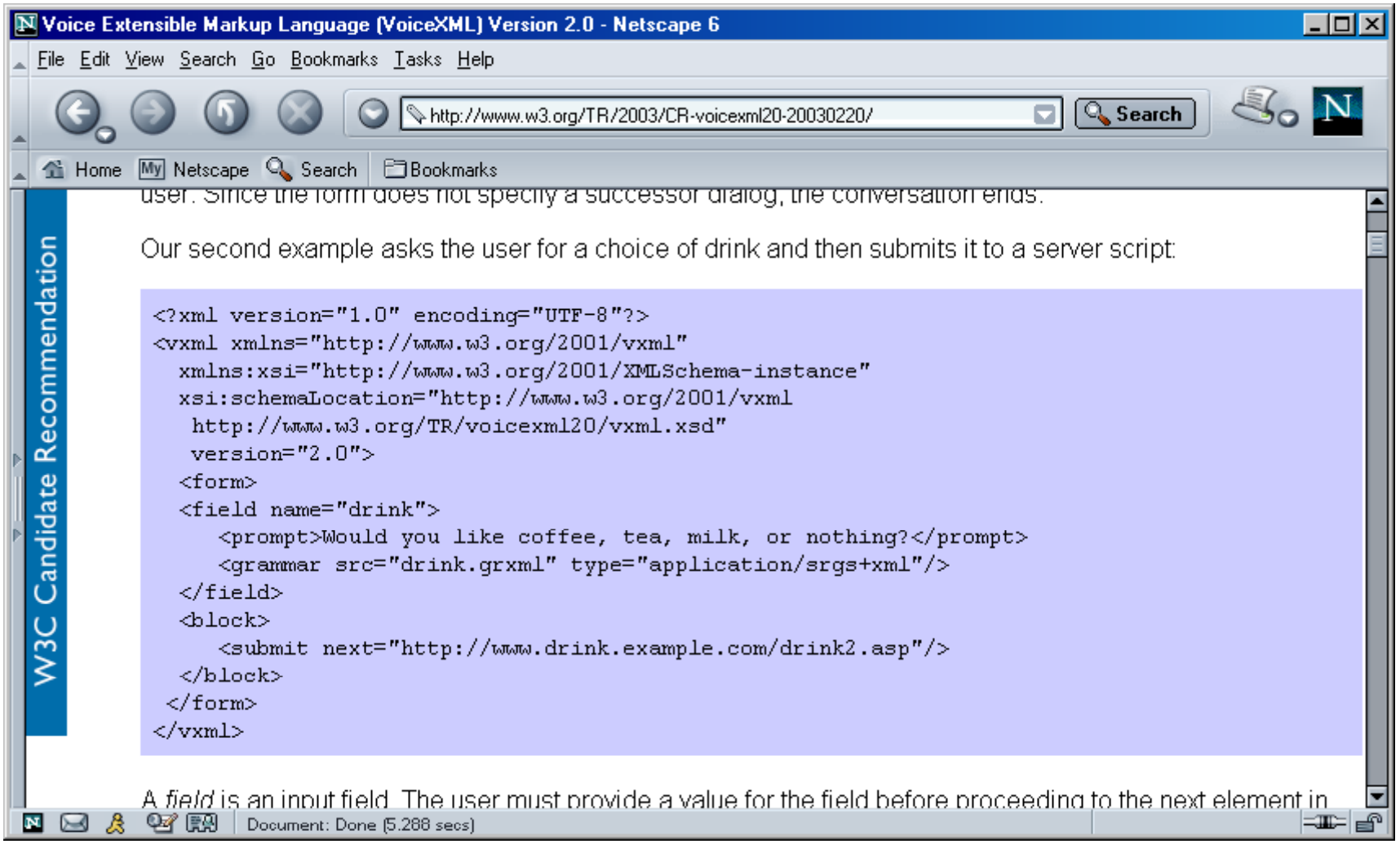
- Spracherkennung: Akustisches + Sprachmodell.

$$\arg\max_{(w_1,\ldots,w_T)} P(w_1,\ldots,w_T \mid Signal)$$

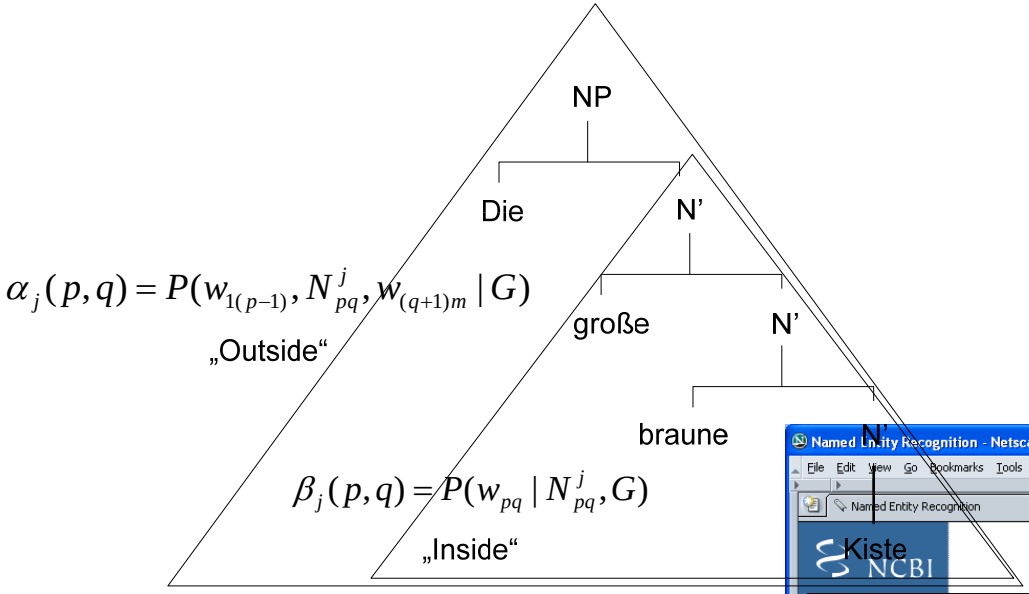
$$= \arg\max_{(w_1,\ldots,w_T)} \underbrace{P(Signal \mid w_1,\ldots,w_T)}_{\text{Akustisches Modell}} \underbrace{P(w_1,\ldots,w_T)}_{\text{Sprachmodell}}$$

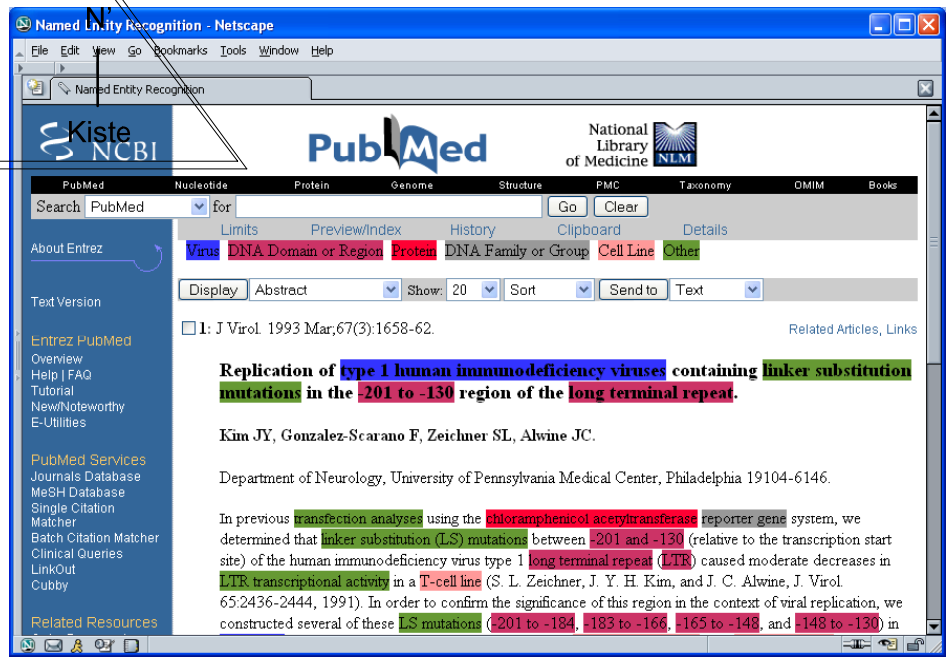


Sprachportale



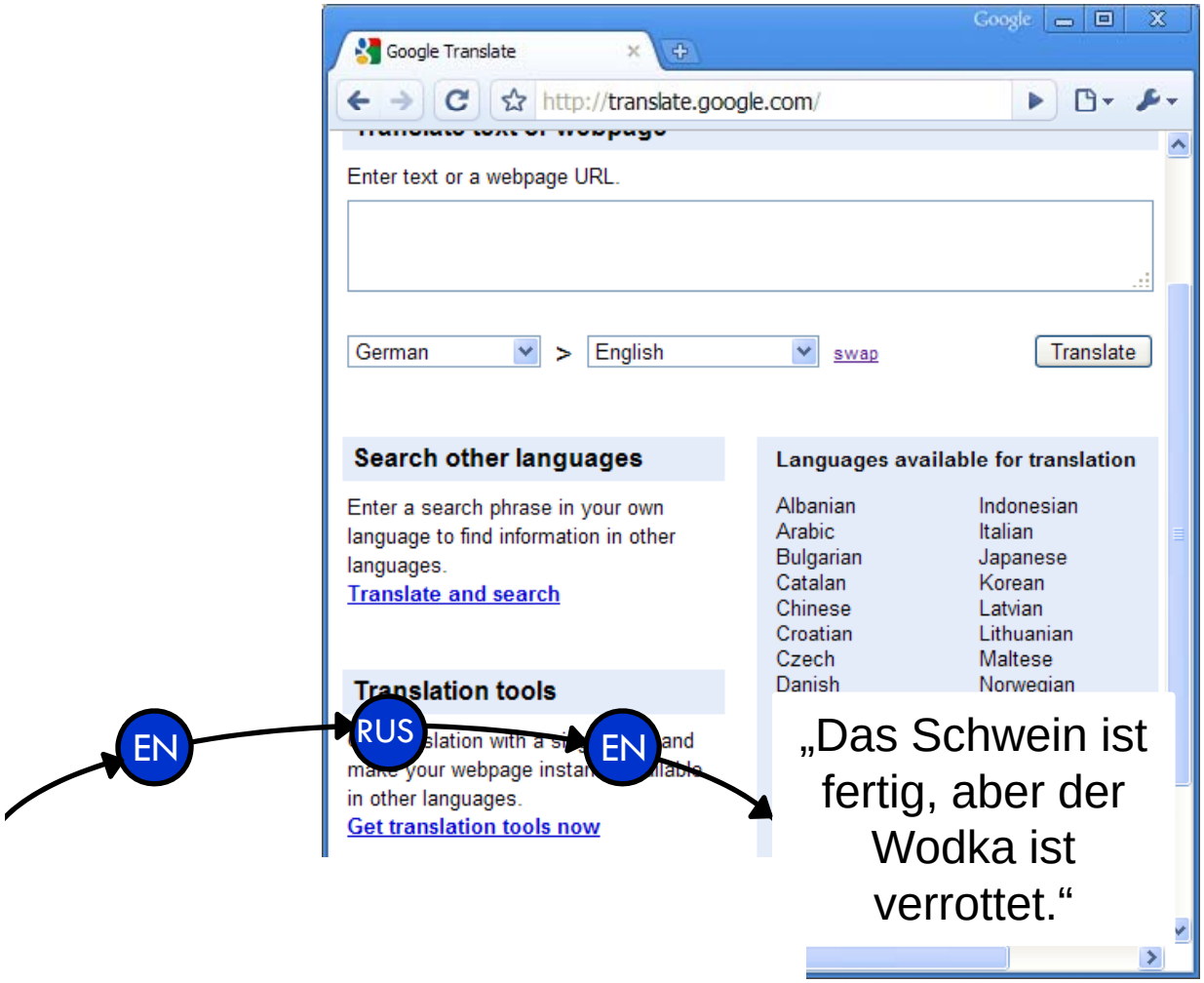
Part-of-Speech Tagging, Named Entity Recognition, Parsing





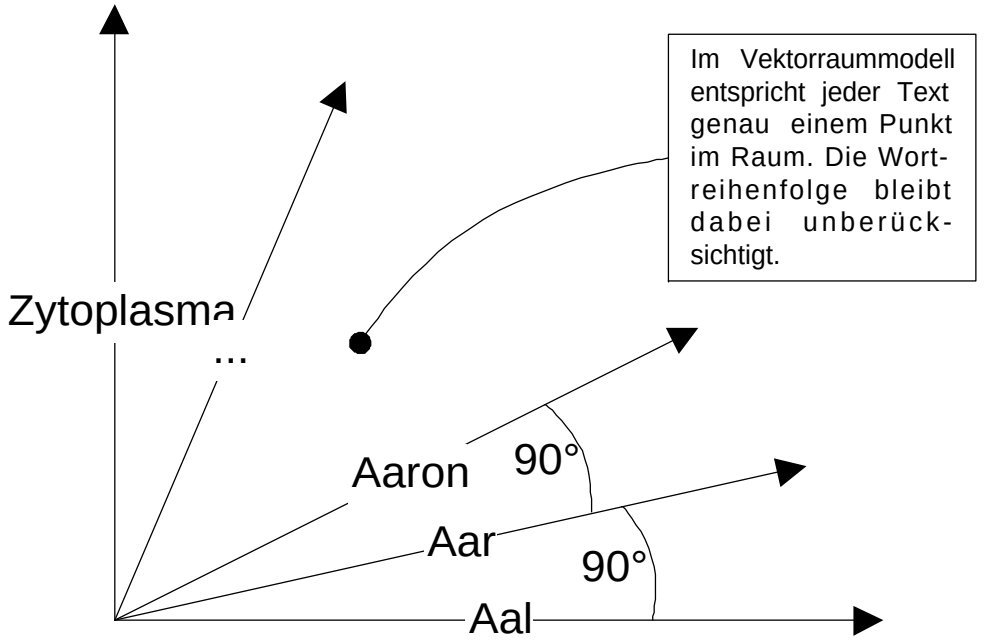
Übersetzung

„Das Fleisch ist willig, aber der Geist ist schwach.“

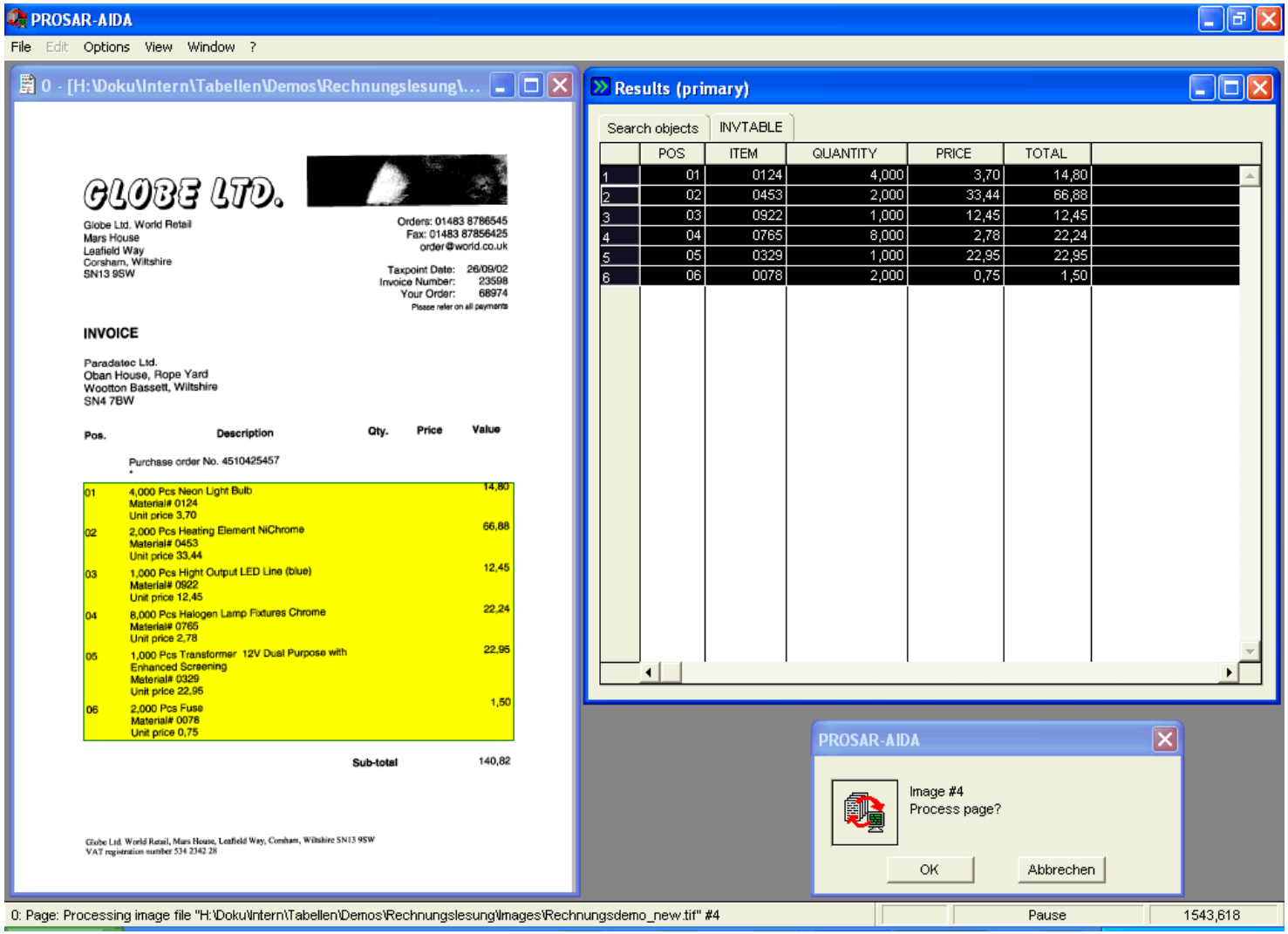


Vektorraummodell

- Repräsentation von Texten.
  - Textklassifikation,
  - Clusteranalyse,
  - Textähnlichkeit,
  - Suche.



Textklassifikation, Informationsextraktion



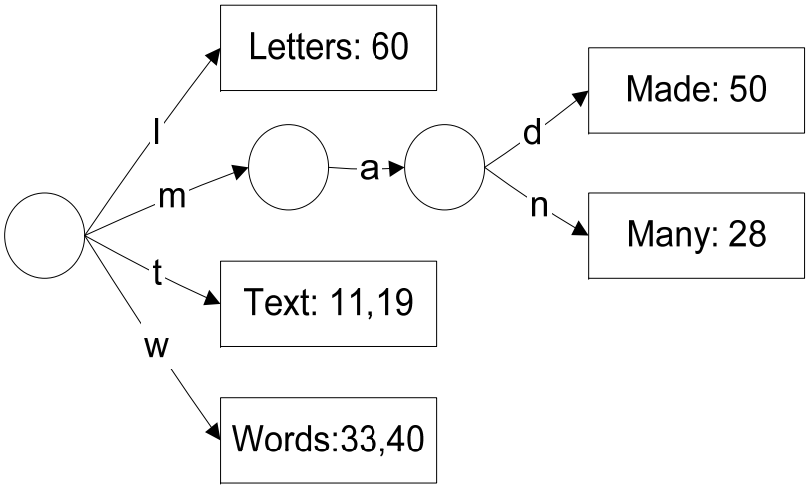
Indexstrukturen

- Schnelle Suche in großen Textsammlungen.

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

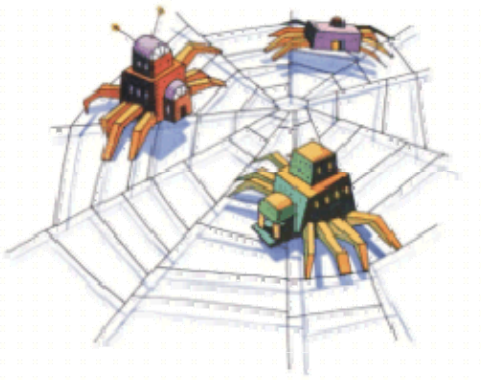


Linkanalyse

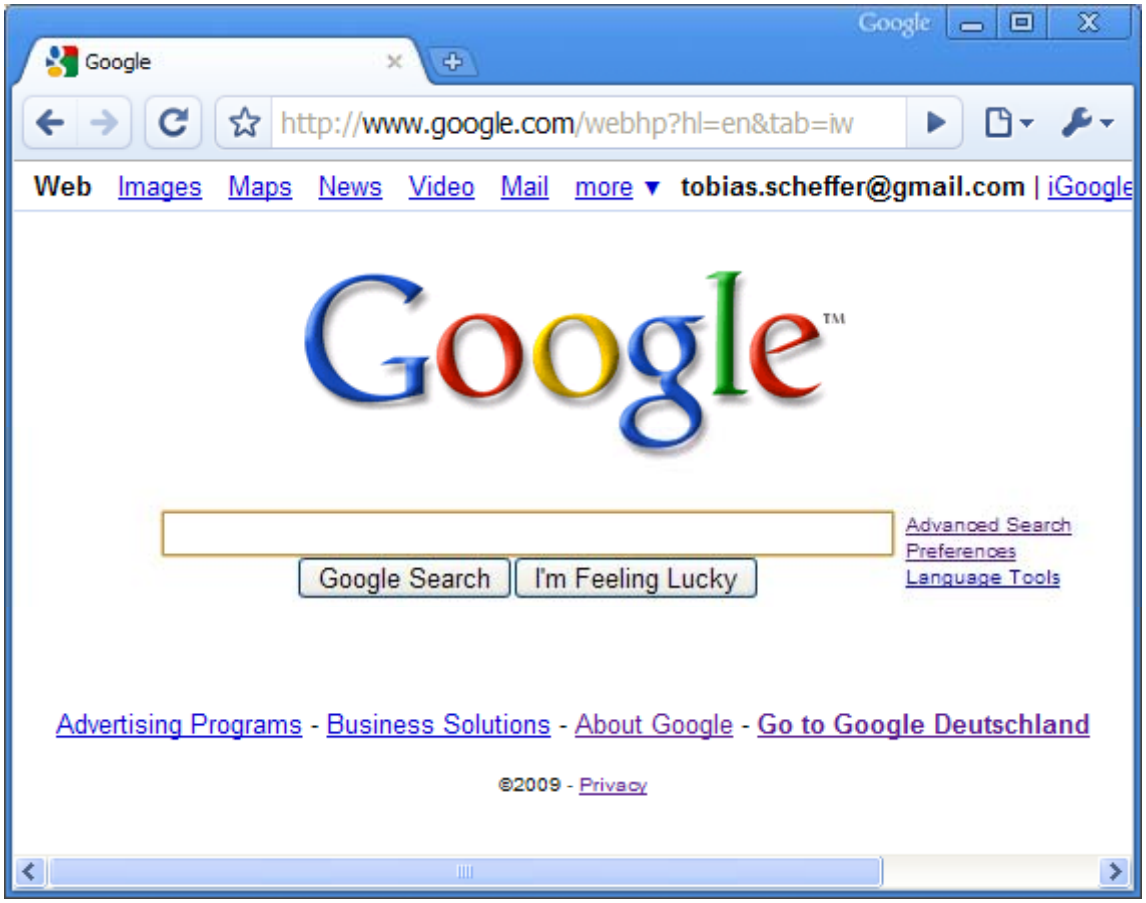
- Relevanz-Ranking: Analyse der Linkstruktur.

Crawling

- Welche URL wann besuchen?
  - Endlos-URLs, dynamische Seiteninhalte.
  - Aktualisierungshäufigkeiten und Zeitpunkte.
  - Identische Seiten.
  - Link-Spam.



Websuche


Google
http://www.google.com/webhp?hl=en&tab=iw
Web Images Maps News Video Mail more ▼ tobias.scheffer@gmail.com | iGoogle
Google™
Google Search
I'm Feeling Lucky
Advanced Search
Preferences
Language Tools
Advertising Programs - Business Solutions - About Google - Go to Google Deutschland
©2009 - Privacy

Fragen?