Maschinelle Sprachverarbeitung
Case Report

- Patient with pneumonia and cough
- Normal dosage of codeine
- Patient not responding any more at day 4
- What’s going on?
  - PubMed „Codeine intoxication“ -> 170 abstracts
  - Aren’t there better ways?

Case report from Univ. Hospital Geneva, thanks to Christian Meisel, Roche
Using AliBaba

Codeine intoxication associated with ultrarapid CYP2D6 metabolism.

Codeine is bioactivated by CYP2D6 into morphine, which then undergoes further glucuronidation.
Finding Relevant Knowledge

- “Find information about …”
- Much knowledge is in text (and only text)
- **Find articles** with information about …
  - PubMed/Medline
  - Information Retrieval
- **Find information** … inside each article
  - Reading many abstracts is tedious
  - Information Extraction
Question

“Which proteins are associated to RAB5?”

Class of terms;
not a term
PubMed Results 2008
What we Need to do

Z-100 is an arabinomannan extracted from Mycobacterium tuberculosis that has various immunomodulatory activities, such as the induction of interleukin 12, interferon gamma (IFN-gamma) and beta-chemokines. The effects of Z-100 on human immunodeficiency virus type 1 (HIV-1) replication in human monocyte-derived macrophages (MDMs) are investigated in this paper. In MDMs, Z-100 markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. Z-100 was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the env gene is defective and the nef gene is replaced with the firefly luciferase gene) when this vector was transfected directly into MDMs. These findings suggest that Z-100 inhibits virus replication, mainly at HIV-1 transcription. However, Z-100 also downregulated expression of the cell surface receptors CD4 and CCR5 in MDMs, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that Z-100 induced IFN-beta production in these cells, resulting in induction of the 16-kDa CCAAT/enhancer binding protein (C/EBP) beta transcription factor that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of p38 mitogen-activated protein kinases (MAPK), indicating that the p38 MAPK signalling pathway was involved in Z-100-induced repression of HIV-1 replication in MDMs. These findings suggest that Z-100 might be a useful immunomodulator for control of HIV-1 infection.
**Find Entities**

**Z-100** is an *arabinomannan* extracted from *Mycobacterium tuberculosis* that has various immunomodulatory activities, such as the induction of *interleukin 12*, *interferon gamma* (*IFN-gamma*) and beta-chemokines. The effects of Z-100 on human immunodeficiency virus type 1 (HIV-1) replication in human *monocyte-derived macrophages* (MDMs) are investigated in this paper. In MDMs, Z-100 markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic *Moloney murine leukemia virus* or *vesicular stomatitis virus G* envelopes. Z-100 was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the *env* gene is defective and the *nef* gene is replaced with the *firefly luciferase* gene) when this vector was transfected directly into MDMs. These findings suggest that Z-100 inhibits virus replication, mainly at HIV-1 transcription. However, Z-100 also downregulated expression of the *cell surface* receptors CD4 and CCR5 in MDMs, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that Z-100 induced *IFN-beta* production in these cells, resulting in induction of the 16-kDa *CCAAT/enhancer binding protein (C/EBP)* beta transcription factor that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of p38 mitogen-activated protein kinases (MAPK), indicating that the p38 MAPK signalling pathway was involved in Z-100-induced repression of HIV-1 replication in MDMs. These findings suggest that Z-100 might be a useful immunomodulator for control of HIV-1 infection.
Z-100 is an arabinomannan extract from Mycobacterium tuberculosis that has various immunomodulatory activities, such as the induction of interleukin 12, interferon gamma (IFN-gamma) and beta-chemokines. The effects of Z-100 on human immunodeficiency virus type 1 (HIV-1) replication in human monocyte-derived macrophages (MDMs) are investigated in this paper. In MDMs, Z-100 markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. Z-100 was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, Z-100 markedly inhibited the expression of the pNL43lucDeltaenv vector (in which the env gene is defective and the nef gene is replaced with the firefly luciferase gene) when this vector was transfected directly into MDMs. These findings suggest that Z-100 inhibits virus replication, mainly at HIV-1 transcription. However, Z-100 also downregulated expression of the cell surface receptors CD4 and CCR5 in MDMs, suggesting some inhibitory effect on HIV-1 entry. These experiments revealed that Z-100 induced IFN-beta production in these cells, which represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of p38 mitogen-activated protein kinases (MAPK), indicating that the p38 MAPK signalling pathway was involved in Z-100-induced repression of HIV-1 replication in MDMs. These findings suggest that Z-100 might be a useful immunomodulator for control of HIV-1 infection.
Detecting Gene Names

*The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.*
Detecting Gene Names (Leser & Hakenberg, 2005)

The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.

- Also: hedgehog, soul, the, white, ...
- State-of-the-art methods reach ~85% in NEN
  - Plus 10% for less stringent boundary definitions
  - Large dicts, CRF, species classification, large background corpus, ...
  - That’s about as high as inter-annotator agreement
- Different performance for other classes (mutations, diseases, functional terms, cell lines, ...)
Chemical Names

[Bux, 2009]: IUPAC-Notation für Valium

7-chloro-1-methyl-5-phenyl-2H-1,4-benzodiazepin-2-one
7-chloro-1-methyl-5-phenyl-3H-1,4-benzodiazepin-2(1H)-one
7-chloro-1-methyl-5-phenyl-1,3-DIHYDRO-2H-1,4-benzodiazepin-2-one
7-chloro-1-methyl-2-oxo-5-phenyl-3H-1,4-benzodiazepine
1-methyl-5-phenyl-7-chloro-1,3-DIHYDRO-2H-1,4-benzodiazepin-2-one
7-chloro-1,3-dihydro-1-methyl-5-phenyl-2H-1,4-benzodiazepin-2-one
7-chloro-1-methyl-5-3H-1,4-benzodiazepin-2(1H)-one

[Klinger et al., 2008]: „Only SMILES and InChl names allow a direct structure search“

InChI=1/C16H13ClN2O/c1-19-14-8-7-12(17)9-13(14)16(18-10-15(19)20)11-5-3-2-4-6-11/h2-9H,10H2,1H3
SMILES=CN1C(=O)CN=C(c2cccc2)c2cc(Cl)c1c12

Trivialnamen, Synonyme, Markennamen, Abkürzungen

Valium = Diazepam = DAP
Drugbank: 117 Brand Names für Diazepam
Typical IE-Workflow

- Information Retrieval
  - Document Retrieval
    - Text Preprocessing
      - Linguistic Annotation
        - Named Entity Recognition
          - Named Entity Normalization
            - Relationship Extraction

- Text Mining
- NLP
Applications in Business Intelligence

• **What problems** are most frequently reported by our customers? Which products, product lines, parts etc.?  
  – Mails, knowledge bases, repair reports, call centers, …

• **How does our customer satisfaction change?**  
  – Tone in communication?  
  – Reports in Blogs, Twitter, …?

• **Can we improve customer self service?**  
  – “Entity Search”  
  – Precise routing and prioritization of requests

Some Recent Students Work

- Can we predict the results of elections using Twitter?
  - Tweet classification, sentiment detection
- What aspects of mobile apps are good / bad?
  - Aspect extraction, topic modelling, sentiment detection
- Can we find texts talking about the biology of stem cells?
  - Text classification, q-gram models
- Can we predict the success of a drug based on papers?
  - Named entity recognition, time series analysis, classification
- Can we semantically cluster tables from the web / papers?
  - Word similarity, text clustering
Modul Maschinelle Sprachverarbeitung

• Vorlesung ~2 SWS
• Übung ~2 SWS
• Slides are English

• Contact
  Ulf Leser
  Raum: IV.401
  Tel: (030) 2093 – 3902
  eMail: leser (...) informatik . hu-berlin . de
Literatur

• Mainly

• Other
  – Weiss, Indurkhya, Zhang, Damerau: „Text Mining“. Springer, 2005
  – Original papers
Anrechnung

Der Kurs (Vorlesung + Praktikum) kann angerechnet werden für:
- Diplom-Informatik, Halle, 8 SP
- Master Informatik, 10 SP

Literatur zur Vorlesung

- Schütze, Manning, Raghavan: „Introduction to Information Retrieval“, MIT Press, 2009

Weitere Literatur und Links

Themen und Termine im Einzelnen

Folien sind hier jeweils nach der Vorlesung als PDF verfügbar. Änderungen möglich. All slides are English, but the course will be held in German.

- Introduction and overview
- Introduction to Information Retrieval
- Evaluation of IR Systems; document normalization
- IR Models I: Boolean, Vector Space, Relevance Feedback
- Exact online substring search: 2-Box and Bayer- Moore
- Indexing terms: Inverted files
- Searching the web: Crawling, PageRank and HITS
- Guest lecture by Prof. Anke Liddle: An Introduction to Linguistics
- Language models
- Wechselwiesen
- Part-of-Speech (POS) tagging
- Collocations and domain-specific terms
- Text classification
- Guest lecture by Dr. Matthias Wendt, Neophon: theo
- Text clustering
- Named Entity Recognition
- Word Sense Disambiguation
- Relationship Extraction
- Abschluss

Weitere Materialien

- Text Retrieval Conference, TREC Reports
Questions?
Questions

- Diplominformatiker?
- Bachelor?
- Semester?

- Special expectations, experiences, questions?
Feedback on Previous Lectures
Content of this Lecture

- A very short primer on Information Retrieval
- A very short primer on Natural Language Processing
- A very short primer on Text Mining
Typical IE-Workflow

Information Retrieval

Document Retrieval

Text Preprocessing

Linguistic Annotation

Named Entity Recognition

Named Entity Normalization

Relationship Extraction

Text Classification

Text Clustering

NLP

Text Mining
Information Retrieval (aka “Search”)

- Find all **documents** which contain the following **words**
- „Leading the user to those documents that will best enable him/her to satisfy his/her need for information“ [Robertson 1981]
  - A user wants to know something
  - The user needs to tell the machine what he wants to know: **query**
  - Posing exact queries is difficult: room for interpretation
  - **Machine interprets query** to compute the (hopefully) best answer
  - Goodness of answer depends on original intention of user, not on the query (relevance)
Difference to Database Queries

- Queries: Formal language versus natural language
- Exactly defined result versus loosely described relevance
- Result set versus ranked list of results
- DB: Posing the right query is completely left to the user
- IR: Understanding the query is a problem of the software
Natural Language Processing

- Making natural language text accessible to a computer
- Multiple levels
  - Find semantic units: words, tokens, phrases, clauses, sentences
    - “Implementing the C4.5 algorithm with languages such as DOT.NET, Java etc. is not as simple as one might think …”
    - “The $\alpha(3)$-helicase-5′ mRNA is …”
  - Find grammatical role of words
  - Find grammatical structure of sentences
  - Find syntactic relationships between entities
- Information may span multiple sentences (co-reference)
- “Understand” the text
„The PAX1 protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.“
Part-Of-Speech Tagging

• Part-of-Speech (POS) is the grammatical class of a word
  - Adverb, verb, adjective, …
  - Verb: Tense, number, …
  - Noun: Gender, case, number, …

  DT  Determiner
  EX  Existential *there*
  FW  Foreign word
  NN  Noun, singular or mass
  NNS Noun, plural
  NNP Proper noun, singular
  NNPS Proper noun, plural
  RP  Particle
  SYM Symbol (math or scientific)
  UH  Interjection
  VB  Verb, base form
  VBD Verb, past tense
  VBG Verb, gerund/present participle
  VBZ Verb, 3rd person/singular, present

• POS tagging: Given a text, assign each word its POS
  - “Does/VBZ flight/NN LH750/NNP serve/VB dinner/NN ?”
  - Caveat: There are different tag sets

• POS tags are very useful for many tasks
  - NER: names of entities should be nouns

• Method: Maximum Entropy, Hidden Markov Models
Parsing

• Revealing the full syntactic structure of a sentence

• Very difficult because of the “ubiquitous’ness” of ambiguities
  – „The fact that Otto knew was surprising“, „Das Fühlen der Hand war abnorm“, „Plakate kleben“

  • Kleben Plakatkleber Plakate an Plakatsäulen oder kleben Plakate durch die Kraft von Plakatkleber an der Säule
Text Mining

- Text Mining = “Data Mining on text”
- Text Mining = “Statistical NLP minus parsing” [Altmann/Schütze]
- Typical tasks
  - Document classification (route emails to the right operator)
  - Document clustering (grouping search results)
  - Information extraction (find all celebrities and their partners)
Example [Juliane Rutsch, 2005]

• Find publications treating the molecular basis of hereditary diseases

• Keyword search generates too many results
  - “Asthma”: 84,884 hits
    • Asthma and cats, factors inducing asthma, treatment, …
  - “Wilson disease”: 4552 hits
    • Including all publications from doctors named Wilson

• Keyword search does not cope with synonyms
Learn what is **typical for a paper** treating the molecular basis of a specific disease

- Here: 25 hereditary diseases, 20 abstracts for each disease

We call this “typical” a **model** of the data

Models are learned from examples using some method

**Classification:** Given a new text, find the model which fits best and **predict the associated class** (disease)

What could we learn from sets of documents?
Wilson's disease, an autosomal recessive disorder, is characterized by the excessive accumulation of copper in the liver. The gene, which encodes a putative copper transporting P-type ATPase, is defective in the patients. To investigate the in vivo function of WND protein as well as its intracellular localization, cDNA was introduced to the Long-Evans Cinnamon rat, known as a rodent model for Wilson's disease, by recombinant adenovirus-mediated gene delivery. An immunofluorescent study and a subcellular fractionation study revealed the transgene expression in liver and its localization to the Golgi apparatus. Moreover, since the synthesis of holoceruloplasmin is disturbed in the Long-Evans Cinnamon rat, the plasma level of holoceruloplasmin, oxidase-active and copper-bound form, was examined to evaluate the function of WND protein with respect to the copper transport. Consequently, the appearance of holoceruloplasmin in plasma was confirmed by Western blot analysis and plasma measurements for the oxidase activity and the copper content. These findings indicate that introduced WND protein may function in the copper transport and that the Golgi apparatus is the likely site for WND protein to manifest its function.

Wilson disease is an autosomal recessive copper transport disorder resulting from defective biliary excretion of copper and subsequent hepatic copper accumulation and liver failure if not treated. The disease is caused by mutations in the ATP7B (WND) gene, which is expressed predominantly in the liver and encodes a copper-transporting P-type ATPase that is structurally and functionally similar to the Menkes protein (MNK), which is defective in the X-linked copper transport disorder Menkes disease. The toxic milk (tx) mouse has a clinical phenotype similar to Wilson disease patients and, recently, the tx mutation within the murine WND homologue (Wnd) of this mouse was identified, establishing it as an animal model for Wilson disease. In this study, cDNA constructs encoding the wild-type (Wnt-wt) and mutant (Wnt-tx) Wilson proteins (Wnd) were generated and expressed in Chinese hamster ovary (CHO) cells. The tx mutation disrupted the copper-induced relocalization of Wnd in CHO cells and abrogated Wnd-mediated copper resistance of transfected CHO cells. In addition, co-localization experiments demonstrated that while Wnd and MNK are located in the Golgi network in basal copper conditions, with elevated copper, these proteins are sorted to different destinations within the same cell. Ultrastructural studies showed that with elevated copper levels, Wnd accumulated in large multi-vesicular structures resembling late endosomes that may represent a novel compartment for copper transport.

We have previously developed a functional assay in yeast for the copper transporter, ATP7B, defective in Wilson's disease (WND). Analysis of WND variant ATP7B proteins revealed that several were able to completely, or nearly completely, complement a mutant yeast strain in which the ATP7B ortholog CCC2 was disrupted, indicating that these ATP7B proteins retained copper transport activity. We analyzed the intracellular localization of these active WND ATP7B variant proteins using transient transfection of Chinese hamster ovary cells and triple-label immunofluorescence microscopy, as a second possible aspect of defective function. Two ATP7B variants, Asp765Asn and Leu776Val, which have normal copper transport activity in yeast, retained partial normal Golgi network localization, but were predominantly mislocalized throughout the cell. Asp765Asn and Leu776Val proteins were capable of only partial copper-dependent redistribution. WND variant protein Arg778Leu, which has defective function in yeast, was extensively mislocalized, presumably to the endoplasmic reticulum. ATP7B variant proteins Gly943Ser, which has nearly normal function in yeast, and CysProCys/Ser (mutation of the conserved CysProCys motif to SerProSer), inactive in yeast, were localized normally but were unable to redistribute in response to copper. Localization data from this study, binned with functional data from our yeast studies, provide a
Vector Space Model (VSM)

• Each document is converted into a vector
  - Dimensions: All different words in all documents
  - Order of words (and hence grammar) is ignored
  - Value of a dimension is binary (word in doc / not) or a count (number of appearances of word in doc) or …

<table>
<thead>
<tr>
<th>- Asthma</th>
<th>5</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>- 5q31-q33</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>- Cri-du-Chat Syndrome</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>- 5p15.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>- Schizophrenia</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>- 1q21-1q22</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>- Diacephapam</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>- Lisino</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>- mutation</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>- gene</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>- disease</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Documents in VSM

Usually, in TM there are \( \sim 100,000 \) and not just 2 dimensions.
A simple Classifier: Nearest-Centroid

Wilsons Disease

Asthma

Phenylketunuria

New document
What if …
OMIM-Datenbank → Diseases and training documents

Training set: 25 diseases, a 15 docs → Preprocessing (Stemming, stop words) → Generate feature vectors → Train the model

Test set: 25 diseases, a 5 doc → Preprocessing (Stemming, stop words) → Generate feature vectors → Classification → Evaluation

Complete Workflow
Results

% 67 66 65 64 63 62 61 60 59 58 57 56 55 54 53 52 51 50
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14

Different configurations

Precision Recall F-Measure
What we will not cover

• Linguistic analysis beyond POS / parsing
• Spoken language
• Machine translation
• Cross-language search / analysis
• User interfaces
• Special classification problems: Sentiment analysis, question answering, Watson
• Topic modelling
• …