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
Codeine intoxication associated with ultrarapid CYP2D6 metabolism.

Codeine is bioactivated by CYP2D6 into morphine, which then undergoes further glucuronidation.
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 (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 envelope) envelopes. Z-100 was also 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. Experiments revealed that Z-100 induced IFN-beta production in these cells, resulting in expression 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.
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 lymphotrophic 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, …)
Typical IE-Workflow

- Information Retrieval
- Document Retrieval
  - Text Preprocessing
  - Linguistic Annotation
    - Named Entity Recognition
    - Named Entity Normalization
    - Relationship Extraction
- NLP
- Text Mining
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
- Can active learning help for finding gene relationships?
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

- **Highly recommended**

- **Other**
  - **Original papers**
Anrechnung
Der Kurs (Vorlesung + Praktikum) kann angerechnet werden für
- Diplom-Informatik, Halbburg, 10 SP
- Master Informatik, 10 SP

Literatur zur Vorlesung
- Schütze, Manning, Rajaraman: "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
- IR Models II: Probabilistic Retrieval, Latent Semantic Indexing
  (Korregierte Version, 20.5.2009)
- 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 Lüdeling: An Introduction to Linguistics
- Language models
- Nouns
- Part-of-Speech (POS) tagging
- Collocations and domain-specific terms
- Text classification
- Guest lecture by Dr. Matthias Wendt: Neophonie: the...
- Text clustering
- Named Entity Recognition
- Word Sense Disambiguation
- Relationship Extraction
- Abitur

Weitere Materialien
- Text Retrieval Conference, TREC homepage
Questions?
Questions

• Diplominformatiker?
• IBI / Wirtschaftsinformatik?
• Bachelor?
• Semester?

• Special expectations, experiences, questions?
Feedback MaschSprach 2015/2016

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>M</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>M</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>M</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>M</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>M</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>M</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>M</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>M</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>M</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>M</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>M</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Konzeption</th>
<th>Verständlichkeit</th>
<th>Übungshilfe</th>
<th>Veränderliche Antworten</th>
<th>Gute Atmosphäre</th>
<th>Tempo</th>
<th>Schwierigkeit</th>
<th>Arbeitsaufwand</th>
<th>Vorbereitung</th>
<th>Abweichung vom Optimum</th>
<th>Abweichung pro Frage</th>
<th>Gelehrt</th>
<th>Warum kommen?</th>
<th>Studiengang</th>
<th>Fachsemester</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,89</td>
<td>5,52</td>
<td>5,26</td>
<td>5,74</td>
<td>5,56</td>
<td>4,72</td>
<td>5,32</td>
<td>5,74</td>
<td>2,84</td>
<td>2,89</td>
<td>2,95</td>
<td>1,21</td>
<td>1,26</td>
<td>0,8</td>
<td>3,7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>M</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

6,40
Was ich ändern will

• NER und RE etwas unsystematisch
• Scalable IE rausnehmen und was dazu?
  – Deep learning and word embeddings
  – Question Answering, Opinion Mining, Topic modelling?
• Vorstellende Gruppen vorher festlegen?
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
Phases in Text Mining

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
  – 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 α(3)-helicase-5’ mRNA is …”
  – Find grammatical role of words
  – Find grammatical structure of sentences
  – Find syntactic relationships between entities
  – Draw semantic inferences from a text
  – …

• “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, ...

- 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

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>Determiner</td>
</tr>
<tr>
<td>EX</td>
<td>Existential <em>there</em></td>
</tr>
<tr>
<td>FW</td>
<td>Foreign word</td>
</tr>
<tr>
<td>NN</td>
<td>Noun, singular or mass</td>
</tr>
<tr>
<td>NNS</td>
<td>Noun, plural</td>
</tr>
<tr>
<td>NNP</td>
<td>Proper noun, singular</td>
</tr>
<tr>
<td>NNPS</td>
<td>Proper noun, plural</td>
</tr>
<tr>
<td>RP</td>
<td>Particle</td>
</tr>
<tr>
<td>SYM</td>
<td>Symbol (math or scientific)</td>
</tr>
<tr>
<td>UH</td>
<td>Interjection</td>
</tr>
<tr>
<td>VB</td>
<td>Verb, base form</td>
</tr>
<tr>
<td>VBD</td>
<td>Verb, past tense</td>
</tr>
<tr>
<td>VBG</td>
<td>Verb, gerund/present participle</td>
</tr>
<tr>
<td>VBZ</td>
<td>Verb, 3rd person/singular, present</td>
</tr>
</tbody>
</table>
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 (group search results by topics)
  - Information extraction (find all celebrities and their partners)
Clinical Entity Recognition for ICD-9 Code Prediction in Clinical Discharge Summaries

Diploma Thesis Presentation
Jonathan Bräuer
02.10.2017
- Clinical data is often stored in **textual form**
- Reports contain valuable information
  - Diseases, symptoms, treatments, drugs, dosage, family history, lab measurements, images/radiology, progression, ...
  - Many of these not available in structured form
- Especially important: **Disease** (symptoms, phenotypes)
  - For accounting
  - For decision support
• **DATE OF ADMISSION:** MM/DD/YYYY
• **DATE OF DISCHARGE:** MM/DD/YYYY
• **DISCHARGE DIAGNOSES:**
  1. Vasovagal syncope, status post fall.
  2. Traumatic arthritis, right knee.
  3. Hypertension.
• **BRIEF HISTORY:** The patient is an (XX)-year-old female with history of previous stroke; hypertension; COPD, stable; renal carcinoma; presenting after a fall and possible syncope. While walking, she accidentally fell to her knees and did hit her head on the ground, near her left eye. Her fall was not observed, but the patient does not profess any loss of consciousness, recalling the entire event. The patient does have a history of previous falls, one of which resulted in a hip fracture. She has had physical therapy and recovered completely from that...
• **DIAGNOSTIC STUDIES:** All x-rays including left foot, right knee, left shoulder and cervical spine showed no acute fractures. The left shoulder did show old healed left humeral head and neck fracture with baseline anterior dislocation. …
• **HOSPITAL COURSE:**
  1. Fall: The patient was admitted and ruled out for syncopal episode. Echocardiogram was normal, and when the patient was able, …
  2. Status post fall with trauma: The patient was unable to walk normally secondary to traumatic injury of her knee, causing significant pain and swelling. Although a scan showed no acute fractures, …
Goals and Methods

• Predict **discharge diagnosis** based on clinical texts
• Approach 1: **Recognize diseases** in text (NER-based approach)

Extract clinical entities → Map to ICD-9-CM → Compare to assigned codes

• Approach 2: **Predict disease** based on (entire, partial) text (classification-based approach)

Extract clinical entities → Transform into vector space → Train classifiers per code → Determine prediction quality
Disease Names: ICD-9

ICD-9-CM Root

001-139 Infectious and parasitic diseases

001-009 Intestinal infect. dis.

001 Cholera

001.0 Cholera due to vibrio cholerae

001.1 ...

002 Typhoid and parat. fevers

002.0 Typhoid fever

002.1 Paratyphoid fever A

003.2 ...

140-239 Neoplasms

010-018 Tuberculosis

019- ...

240- ...

...
Medical NER Tools Evaluated

- UMLSSpecialist
- UIMA
- OpenNLP
- GATE
- BANNER
- MetaMap
- cTAKES
- HITEx
- NCBO
- DNorm
- UMLSS
- SNOMED
- Any Ontology
- MEDIC
Number of Extracted Concepts (Per Document)
Issues (Typical)

- **Hierarchical classification** - which level of ICD-9?
  - Higher levels: More training data, few classes, high accuracy
    But: Little value
  - Lower levels: Little training data, many classes, low accuracy
    But: High value

- **Mapping** between ontologies
  - Concepts with different syntax & synonyms
  - Concepts at different granularities
  - Conflicting subsumption relationships
  - Diverging coverage
  - ...

- DNorm
- MetaMap
- HITEx
- cTAKES

- Other Ontologies
- UMLS (CUI)
- SNOMED-CT (Code)
- ICD-9-CM
# Results / Evaluation

- 50k discharge summaries
- 7k classes (diagnosis codes)

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>MANU</td>
<td>0.173</td>
<td>0.193</td>
</tr>
<tr>
<td>CTAKES</td>
<td>0.179</td>
<td>0.24</td>
</tr>
<tr>
<td>DNorm</td>
<td>0.092</td>
<td>0.203</td>
</tr>
<tr>
<td>Hitex</td>
<td>0.186</td>
<td>0.254</td>
</tr>
<tr>
<td>Meta</td>
<td>0.19</td>
<td>0.251</td>
</tr>
<tr>
<td>Ncbo</td>
<td>0.174</td>
<td>0.235</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>0.71</td>
<td>0.62</td>
</tr>
<tr>
<td>Ctakes</td>
<td>0.74</td>
<td>0.62</td>
</tr>
<tr>
<td>DNorm</td>
<td>0.53</td>
<td>0.62</td>
</tr>
<tr>
<td>Hitex</td>
<td>0.44</td>
<td>0.52</td>
</tr>
<tr>
<td>Meta</td>
<td>0.54</td>
<td>0.6</td>
</tr>
<tr>
<td>Ncbo</td>
<td>0.51</td>
<td>0.57</td>
</tr>
</tbody>
</table>

For evaluation, 50k discharge summaries and 7k classes (diagnosis codes) were used.
Results / Evaluation

- Baseline: 10 k top concepts 7 k
- Train/test split 90% / 10%
Error Analysis

- False positive code assignments
  - Mapping errors
  - Contextual errors
  - Negation / temporal status

- False negative code assignments
  - Obvious codes not tagged in gold standard (hypertension)
  - Heavy use of abbreviations and acronyms
  - Missing sections
  - Missing mention
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
- ...

Ulf Leser: Maschinelle Sprachverarbeitung