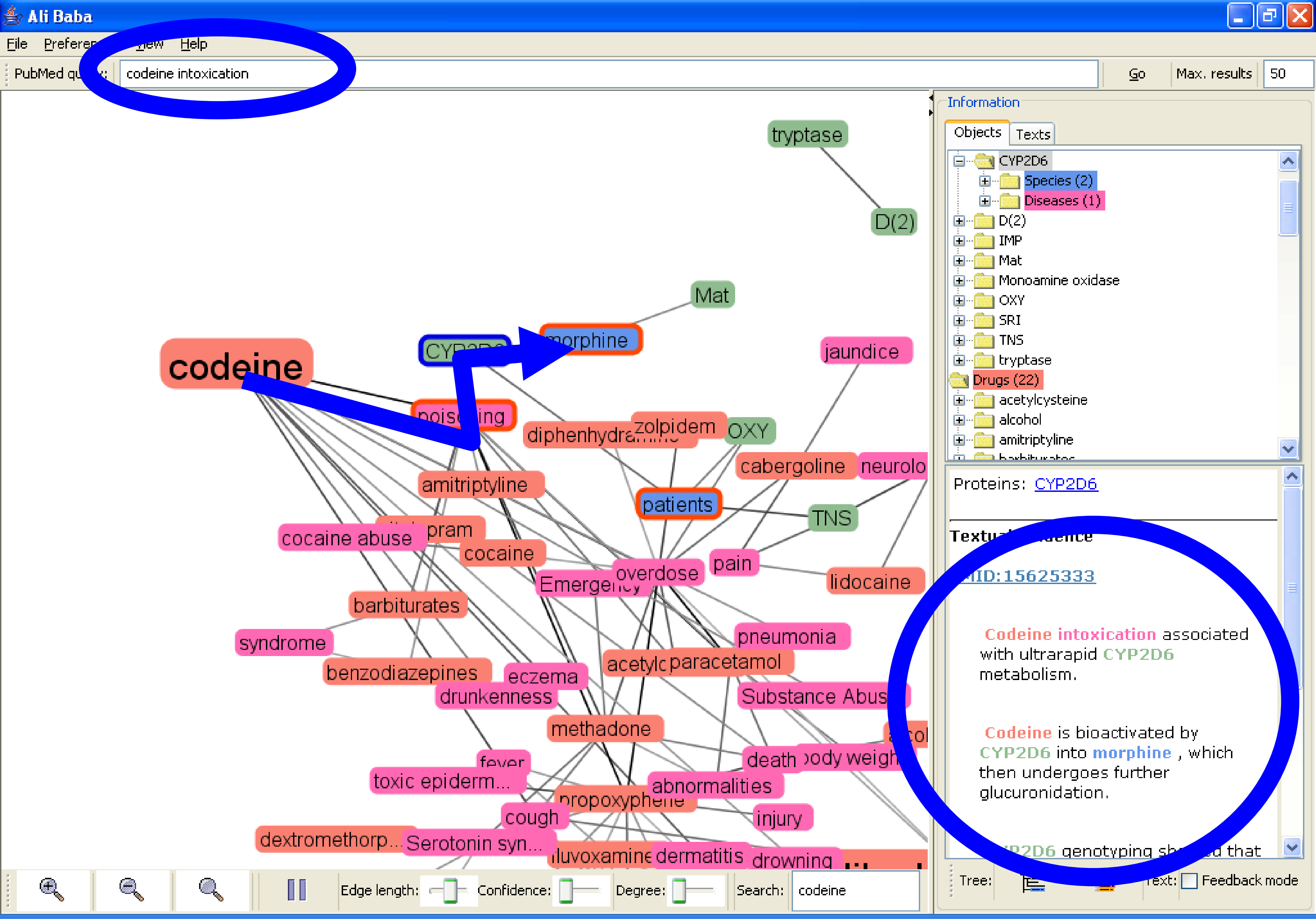
Maschinelle Sprachverarbeitung

Ulf Leser

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


Ali Baba
File
Help
codeine intoxication
Go
Max. results
50
Information
Objects
Texts
CYP2D6
Species (2)
Diseases (1)
D(2)
IMP
Mat
Monoamine oxidase
OXY
SRI
TNS
tryptase
Drugs (22)
acetylcysteine
alcohol
amitriptyline
Proteins: CYP2D6
Codeine intoxication associated with ultrarapid CYP2D6 metabolism.
Codeine is bioactivated by CYP2D6 into morphine , which then undergoes further glucuronidation.
codeine
tryptase
D(2)
Mat
morphine
jaundice
diphenhydramine
zolpidem
OXY
cabergoline
amitriptyline
patients
TNS
cocaine abuse
cocaine
pain
overdose
lidocaine
barbiturates
syndrome
pneumonia
paracetamol
benzodiazepines
eczema
drunkenness
Substance Abus
methadone
death
toxic epiderm...
abnormalities
cough
injury
dextromethorp...
Serotonin syn...
dermatitis
drowning
Edge length:
Confidence:
Degree:
Search:
Tree:
Feedback mode

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.

Find Relationships

Z-100 i n *arabinomann* **induction** rom Mycobacterium tuberculosis that has various immunomodulatory activ he 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 amphot ey murine leukemia virus or vesicular stomatitis virus G envelope . *Z-100* was **inhibit** HIV-1 expression, even when added 24 h after infection. In addition, lly 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 **induces** eriments revealed that *Z-100* induced **IFN-beta** production in these ce tion 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

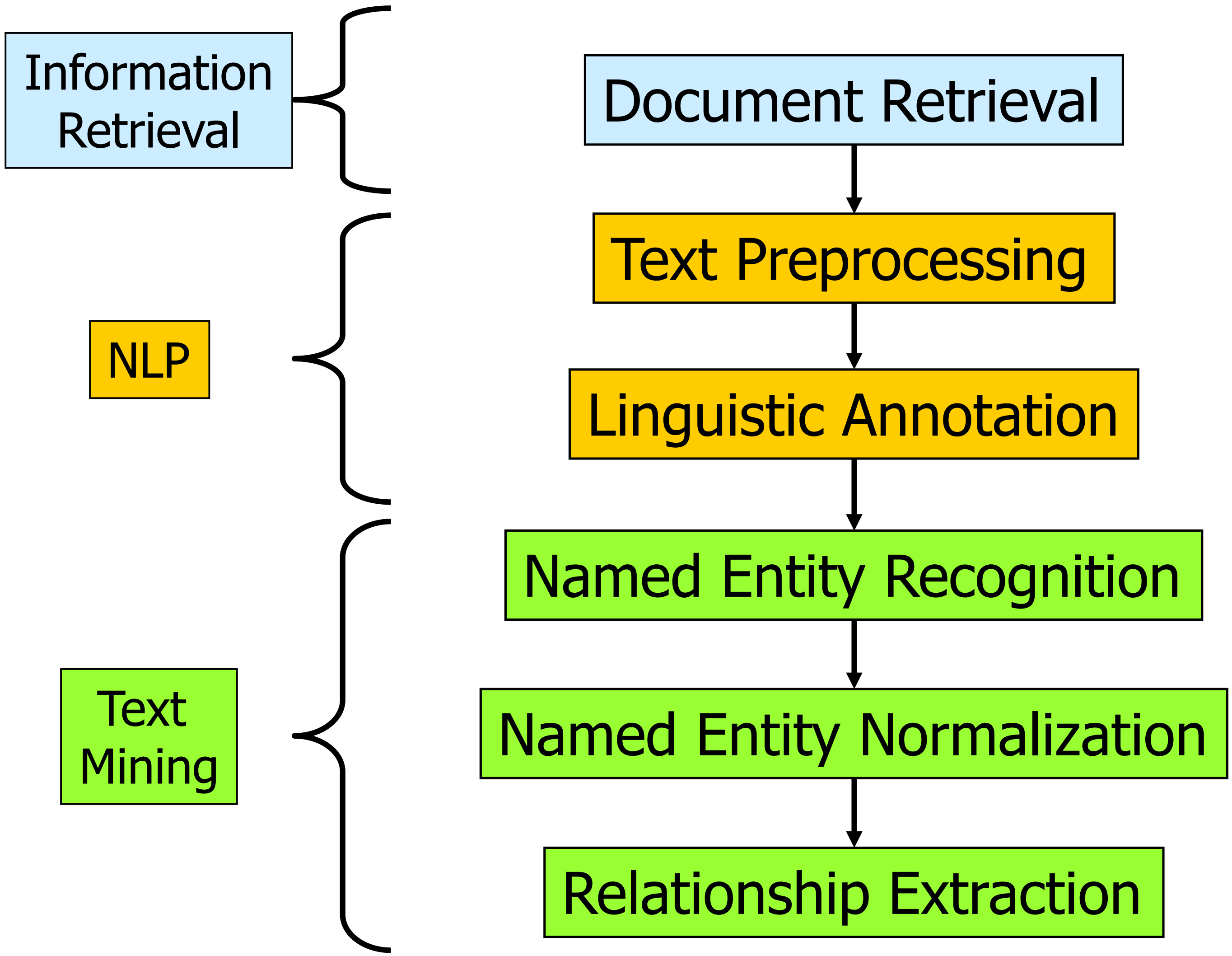
The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.

Detecting Gene Names (Leser & Hakenberg, 2005)

The human T cell leukemia 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, …)

Typical IE-Workflow



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

- See, e.g., Lang, A. and Reinwald, B. (2008). "Nutzung unstrukturierter Daten für Business Intelligence." *Datenbank Spektrum* **25**.

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
  - Starts tomorrow, 18.10.2018
- Slides are English

- Contact

  Ulf Leser

  Raum: IV.401

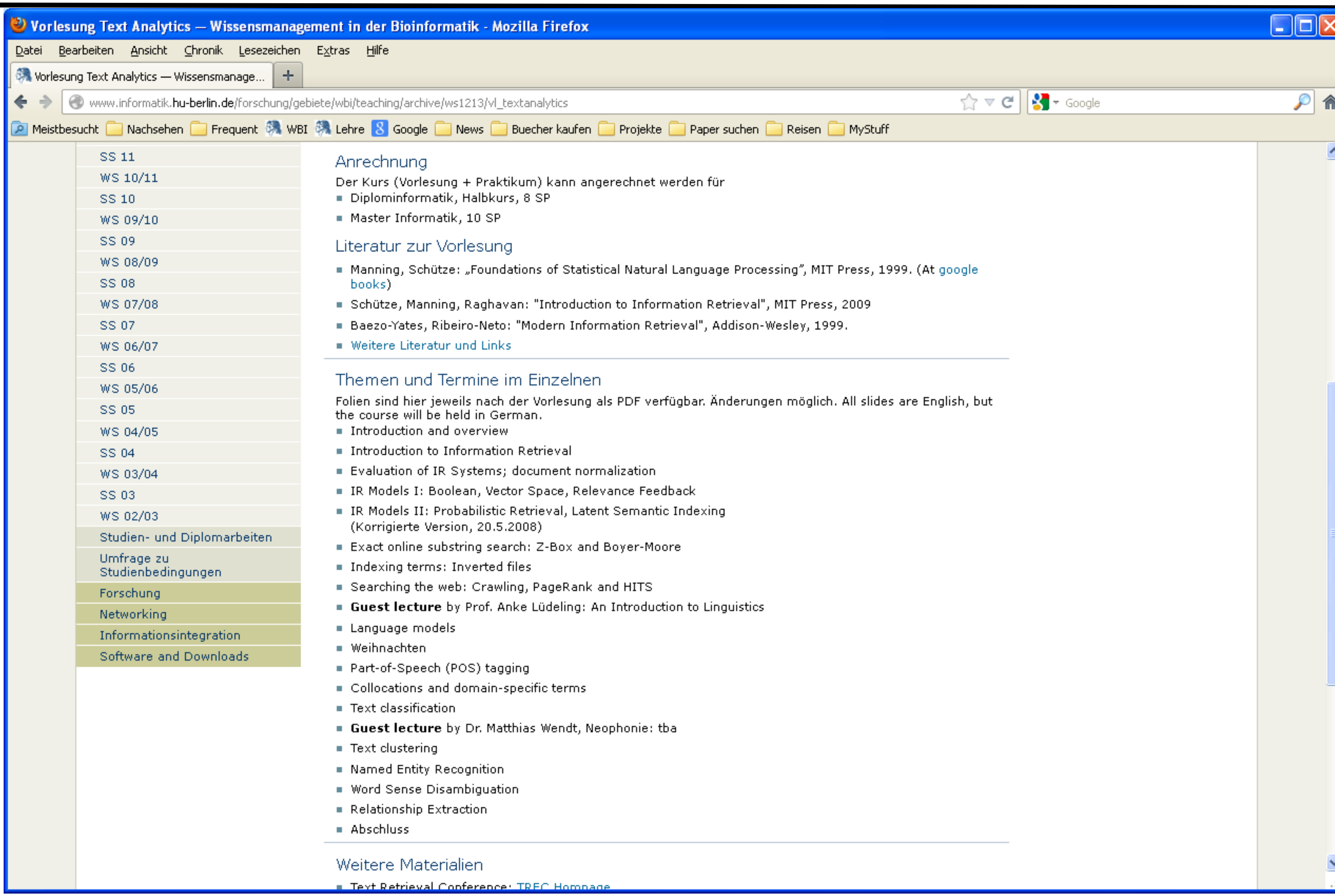
  Tel: (030) 2093 – 3902

  eMail: leser (..) informatik . hu-berlin . de

Literatur

- Highly recommended
  - Manning, C.D., Schütze, H. (1999). „Foundations of Statistical Natural Language Processing", MIT Press.
- Other
  - Weiss, Indurkhya, Zhang: „Fundamentals of Predictive Text Mining", Springer, 2016
  - Aggarwal, Zhai: „ Mining Text Data", Springer, 2012
  - Manning, C. D., Raghavan, P. and Schütze, H. (2008). "Introduction to Information Retrieval", Cambridge University Press.
  - Lemnitzer, L. and Zinsmeister, H. (2010). "Korpuslinguistik - Eine Einführung", narr Studienbücher.
  - Original papers

Web



Vorlesung Text Analytics — Wissensmanagement in der Bioinformatik - Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

Vorlesung Text Analytics — Wissensmanage...

www.informatik.hu-berlin.de/forschung/gebiete/wbi/teaching/archive/ws1213/vl_textanalytics

Google

Meistbesucht Nachsehen Frequent WBI Lehre Google News Buecher kaufen Projekte Paper suchen Reisen MyStuff

- SS 11
- WS 10/11
- SS 10
- WS 09/10
- SS 09
- WS 08/09
- SS 08
- WS 07/08
- SS 07
- WS 06/07
- SS 06
- WS 05/06
- SS 05
- WS 04/05
- SS 04
- WS 03/04
- SS 03
- WS 02/03
- Studien- und Diplomarbeiten
- Umfrage zu Studienbedingungen
- Forschung
- Networking
- Informationsintegration
- Software and Downloads

Anrechnung

Der Kurs (Vorlesung + Praktikum) kann angerechnet werden für

- Diplominformatik, Halbkurs, 8 SP
- Master Informatik, 10 SP

Literatur zur Vorlesung

- Manning, Schütze: „Foundations of Statistical Natural Language Processing", MIT Press, 1999. (At google books)
- Schütze, Manning, Raghavan: "Introduction to Information Retrieval", MIT Press, 2009
- Baezo-Yates, Ribeiro-Neto: "Modern Information Retrieval", Addison-Wesley, 1999.
- 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 (Korrigierte Version, 20.5.2008)
- Exact online substring search: Z-Box and Boyer-Moore
- Indexing terms: Inverted files
- Searching the web: Crawling, PageRank and HITS
- **Guest lecture** by Prof. Anke Lüdeling: An Introduction to Linguistics
- Language models
- Weihnachten
- Part-of-Speech (POS) tagging
- Collocations and domain-specific terms
- Text classification
- **Guest lecture** by Dr. Matthias Wendt, Neophonie: tba
- Text clustering
- Named Entity Recognition
- Word Sense Disambiguation
- Relationship Extraction
- Abschluss

Weitere Materialien

- Text Retrieval Conference: TREC Homepage

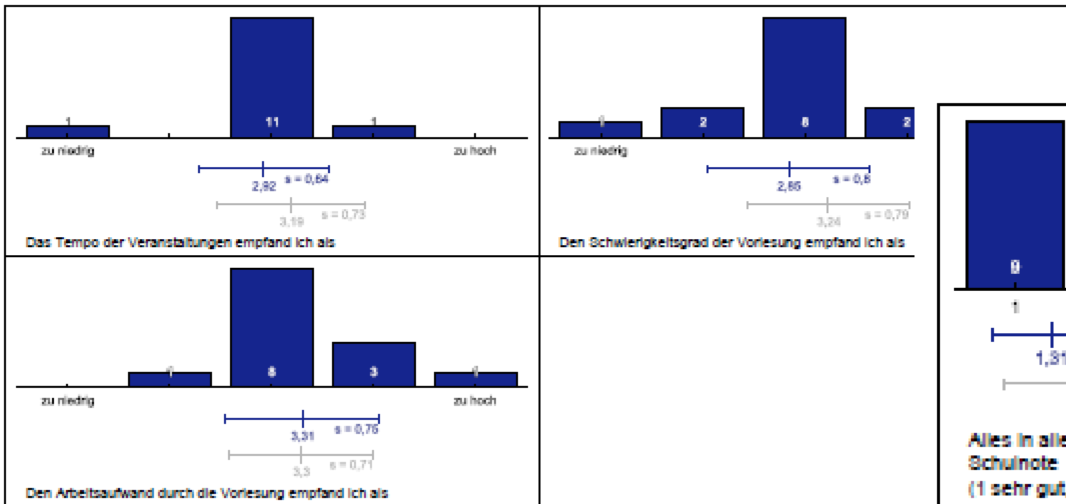
Questions?

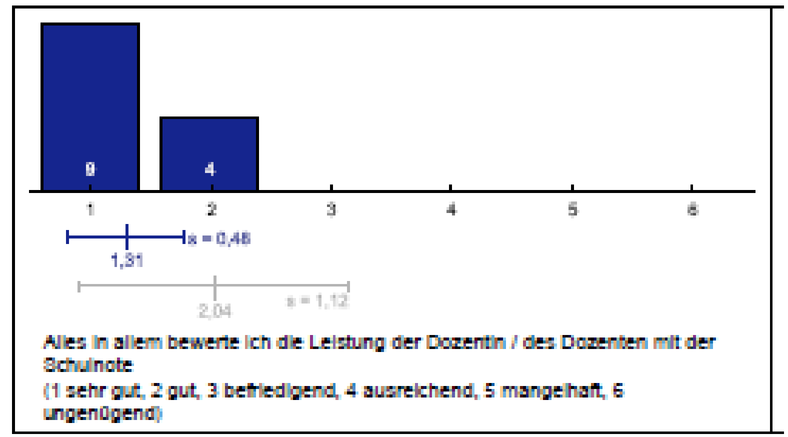
Questions

- Diplominformatiker?
- IBI / Wirtschaftsinformatik?
- Bachelor?
- Semester?

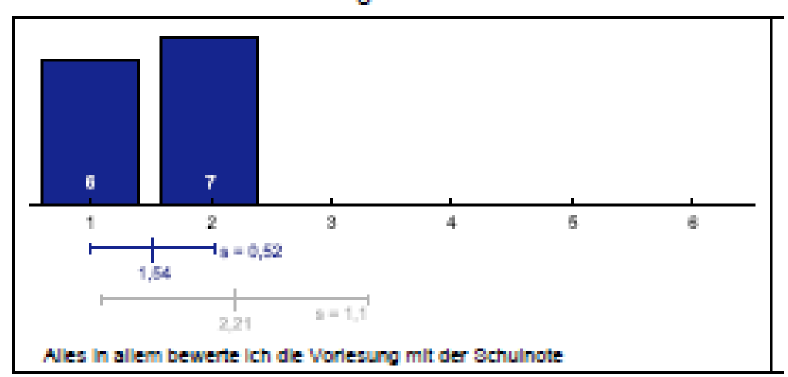
- Special expectations, experiences, questions?

Feedback MaschSprach 2017/2018 (n=13)





Gesamteindruck -Vorlesung



Free comments

Das hat mir gefallen: (DozentInnenfrage)

- 1) Die Übungen haben geholfen den Stoff (vorallem Probleme und Fallstricke) zu verdeutlichen

  2) Die Beispiele waren anschaulich und nicht zuu einseitig (auch wenn der Trend zur Bioinformatik logischerweise erkennbar ist)

  3) Die Struktur der Vorlesung war nachvollziehbar und deutlich
- Beispiele aus der Praxis
- Die vielen Beispiele, zB zu Anwendungen außerhalb von Forschung
- Entspannte Atmosphäre, perfekter Umgang mit den Studenten, Auflockerung durch andere Gesprächsthemen, super Beispiele mit denen man die Inhalte auf Anhieb versteht
- Eine perfekte Vorlesung!
  - sehr ausführliche, gut nachvollziehbare Folien
  - wirklich anwendungsorientiert(case studies, Übung, keine hohe Mathematik)
  - gute Koordination mit der Übung
- sehr interessante Beispiele aus der Bioinformatik, Stoff wurde verständlich erklärt, Fragen konnten jederzeit geäußert werden
- - Viele Beispiele
  - Vortragsstil gut

Welche konstruktiven Anregungen und Verbesserungsvorschläge haben Sie? (DozentInnenfrage)

- 1) Ich verstehe nicht warum die Vorlesung auf deutsch aber die Folien auf englisch sind. Ich würde das einheitlich besser finden.

  2) Einige Algorithmen (z.B. Viterbi) waren für mich nur durch die Vorlesung schwer zu verstehen, es hätte mir geholfen nocheinmal Referenzen zu Quellen zu haben, die das ggf. anders erklären. (Ich weiß das ist auch Eigenverantwortung des Studierenden yada yada. Aber mir hätte es geholfen).

  3) Ich war mir manchmal nicht sicher ob meine Frage angemessen wäre oder ob ich das 'wissen müsste' und hab deshalb dann gar nicht gefragt. Ich weiß nicht wie man diesem Problem vorbeugen kann aber es ist eine Denkanregung ;)
- Der Umfang des Stoffes sowie die zeitintensiven Übungsaufgaben spiegeln IMHO nicht die ausgeschriebenen 5 LP wider
- Es könnten Bezüge zu aktuelleren Verfahren und Techniken hergestellt werden; ich hatte das Gefühl, dass die besprochenen Themen und Methodiken bereits etwas veraltet sind.
- Folien sind okay, könnten aber strukturierter sein.
  Spannnende Einblicke in außerfachliche Themen.
  gerne mehr praktische Beispiele nennen.
- Klarere Beispiele zu den Konzepten auf den Folien. Die Nacharbeitung des Stoffes fand ich manchmal schwer, weil die Beispiele recht minimalistisch waren
- Teilweise wird zu wenig auf die konkrete Implementation von bspw. Algorithmen eingegangen, und zu viel Zeit für die Besprechung von potentiellen Problemen aufgewandt.

Was ich ändern wollte

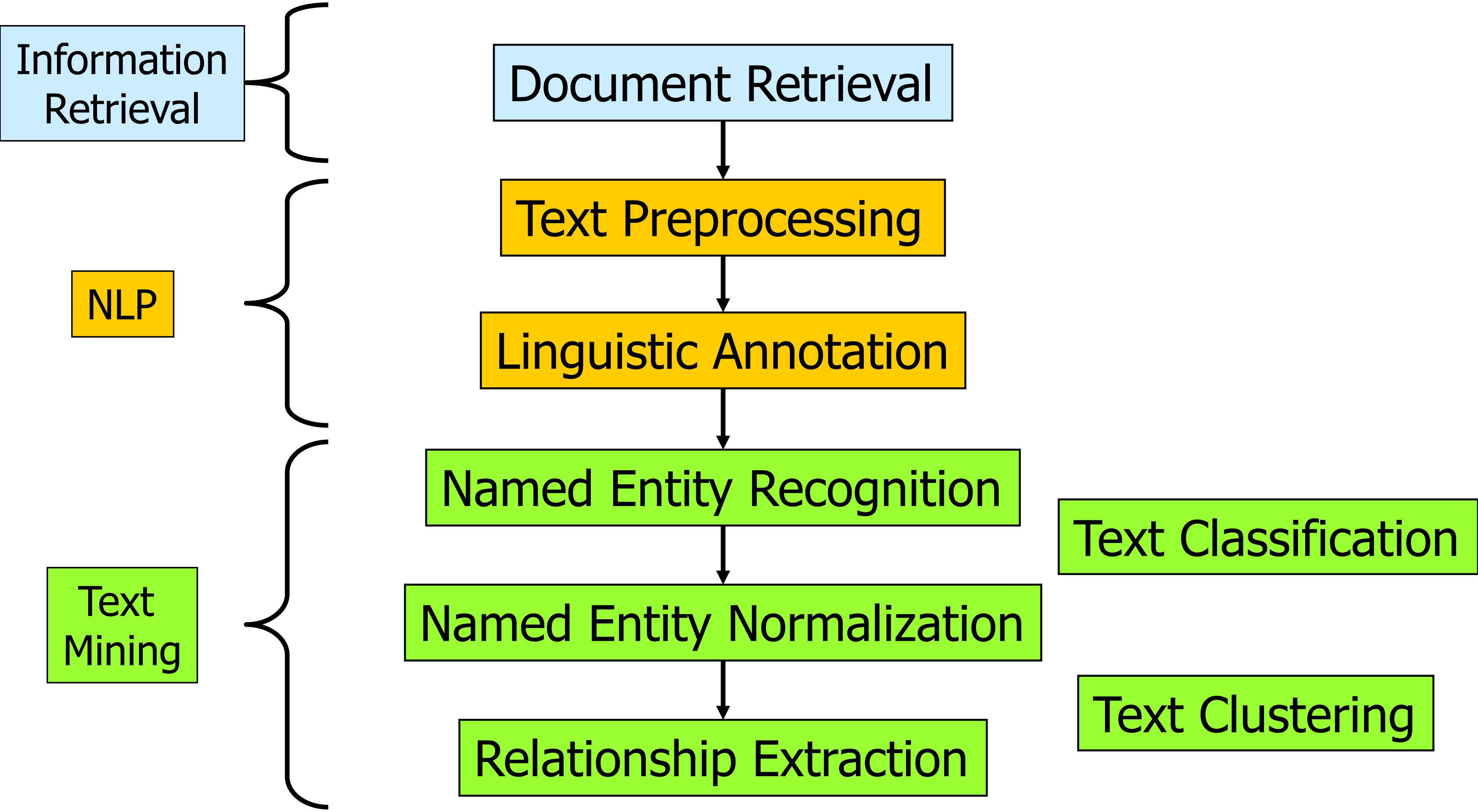
- Rausnehmen: Scalable IE
- Reinnehmen: Deep learning and word embeddings

- Was ich nicht gemacht habe: Auf 3+2 ausbauen
  - Question Answering, Opinion Mining, Topic modelling, …

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 (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

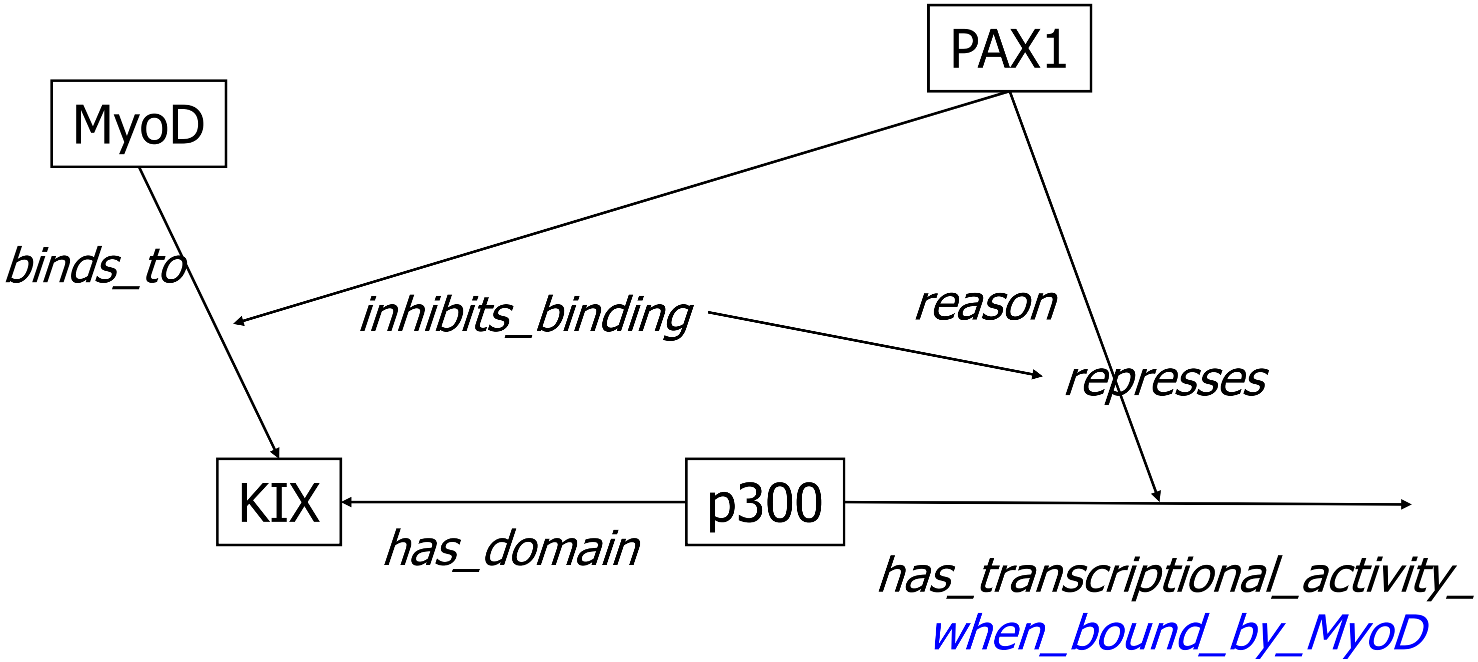
- Queries: Formal language versus natural language
- Exactly defined result versus loosely described relevance
- Result set versus ranked list of results
- DB: Posing the right query is completely left to the user
- IR: Understanding the query is a problem of the software

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

Understanding Text is Difficult (even for us)

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

- 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
- Methods: Maximum Entropy, Hidden Markov Models

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)
  - Question Answering (what will be the weather tomorrow?)

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.
  6. History of chronic obstructive pulmonary disease.
- **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

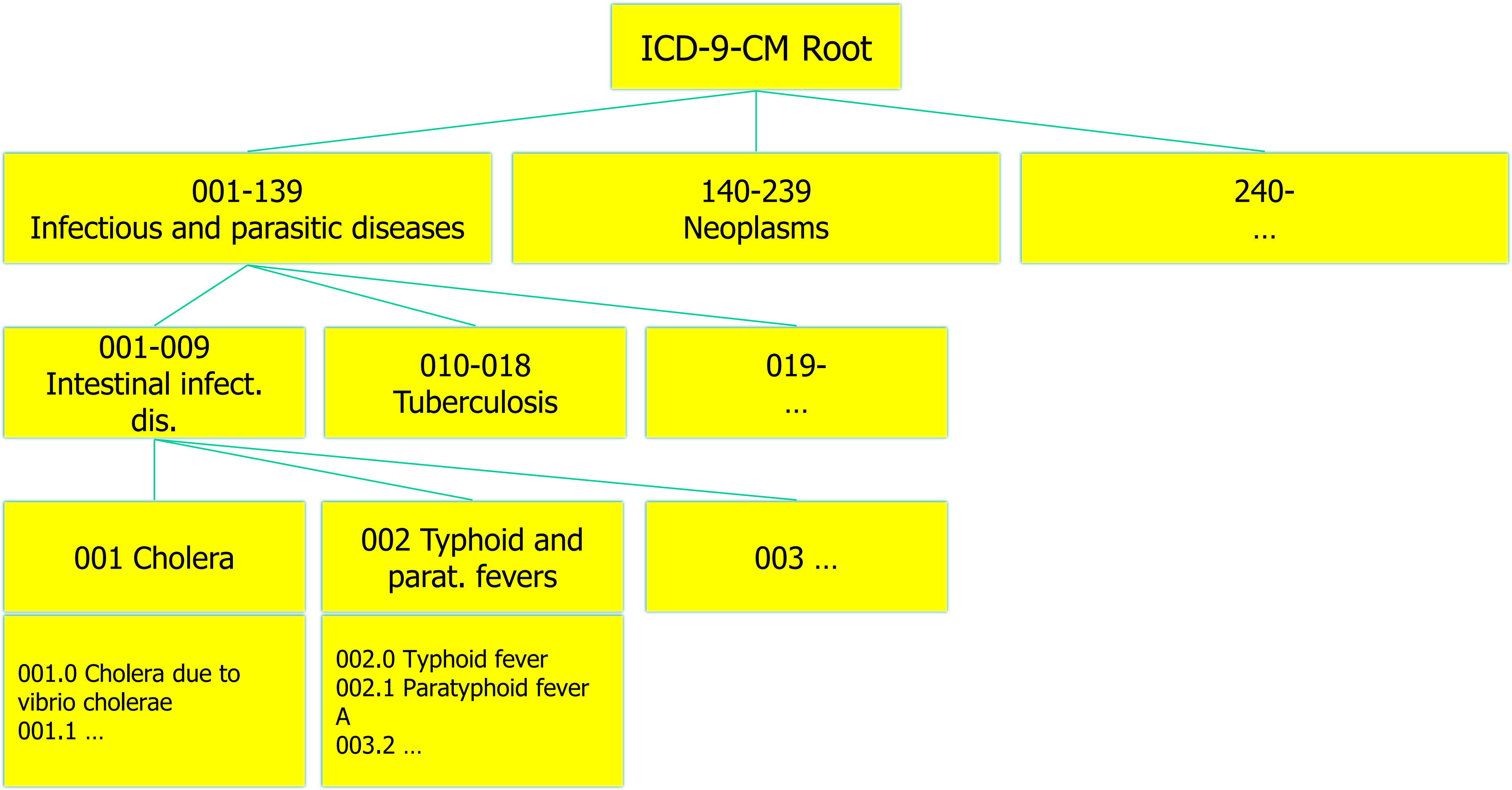
- Predict discharge diagnosis based on clinical texts
- Approach 1: Recognize diseases in text (NER-based approach)

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

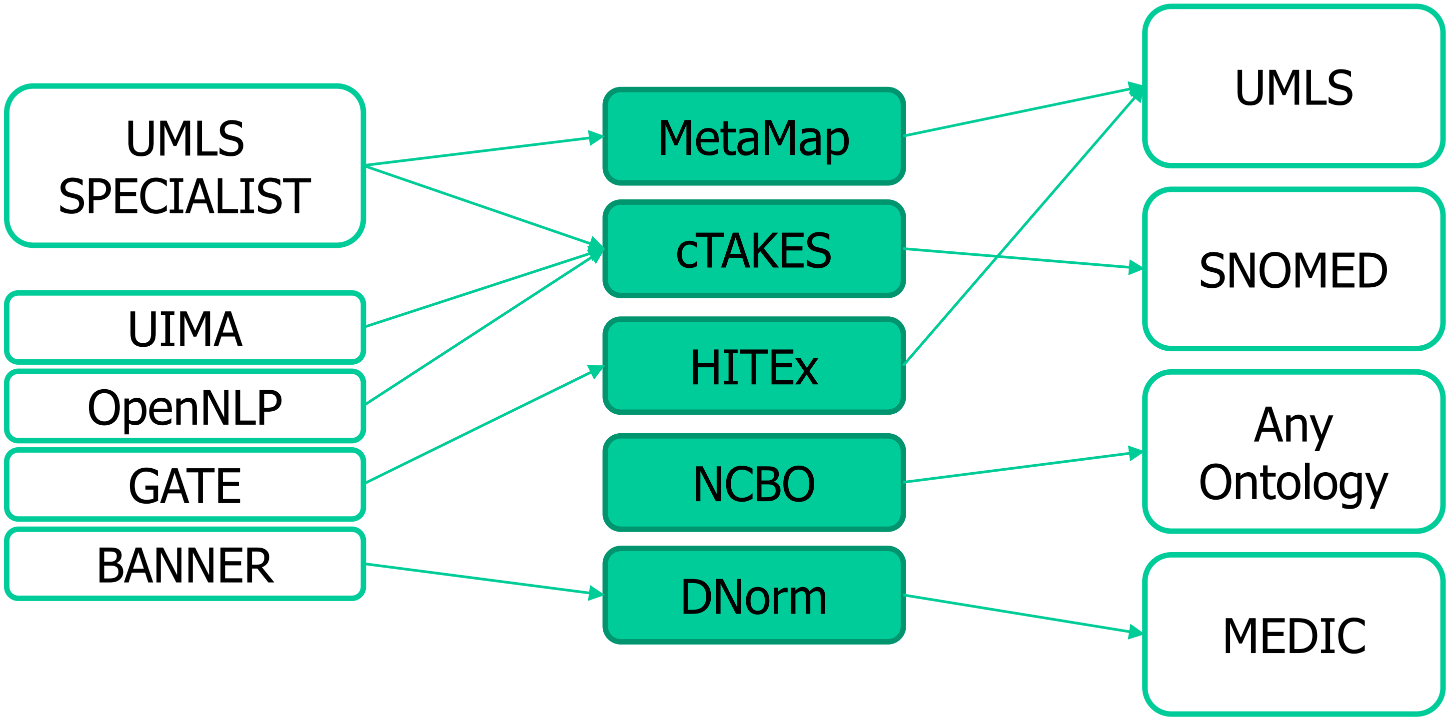
- Approach 2: Predict disease based on (entire, partial) text (classification-based approach)

Extract clinical entities → Transform into vector space → Train classifiers per code → Determine prediction quality

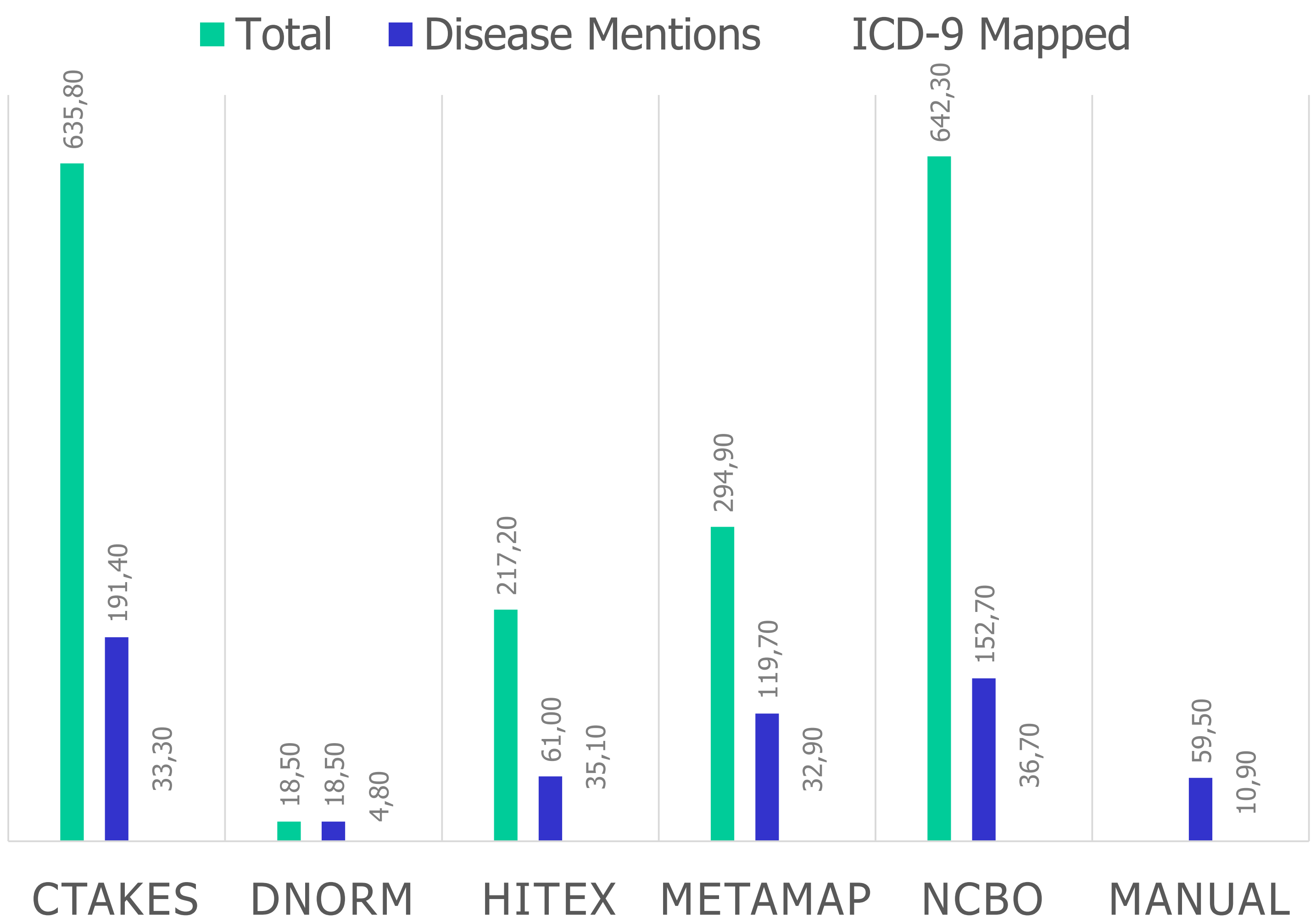
Disease Names: ICD-9



Medical NER Tools Evaluated

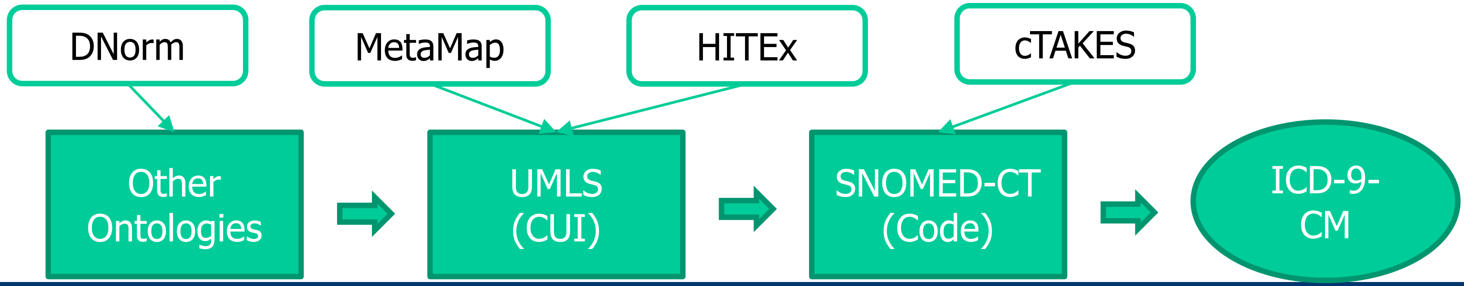


Number of Extracted Concepts (Per Document)



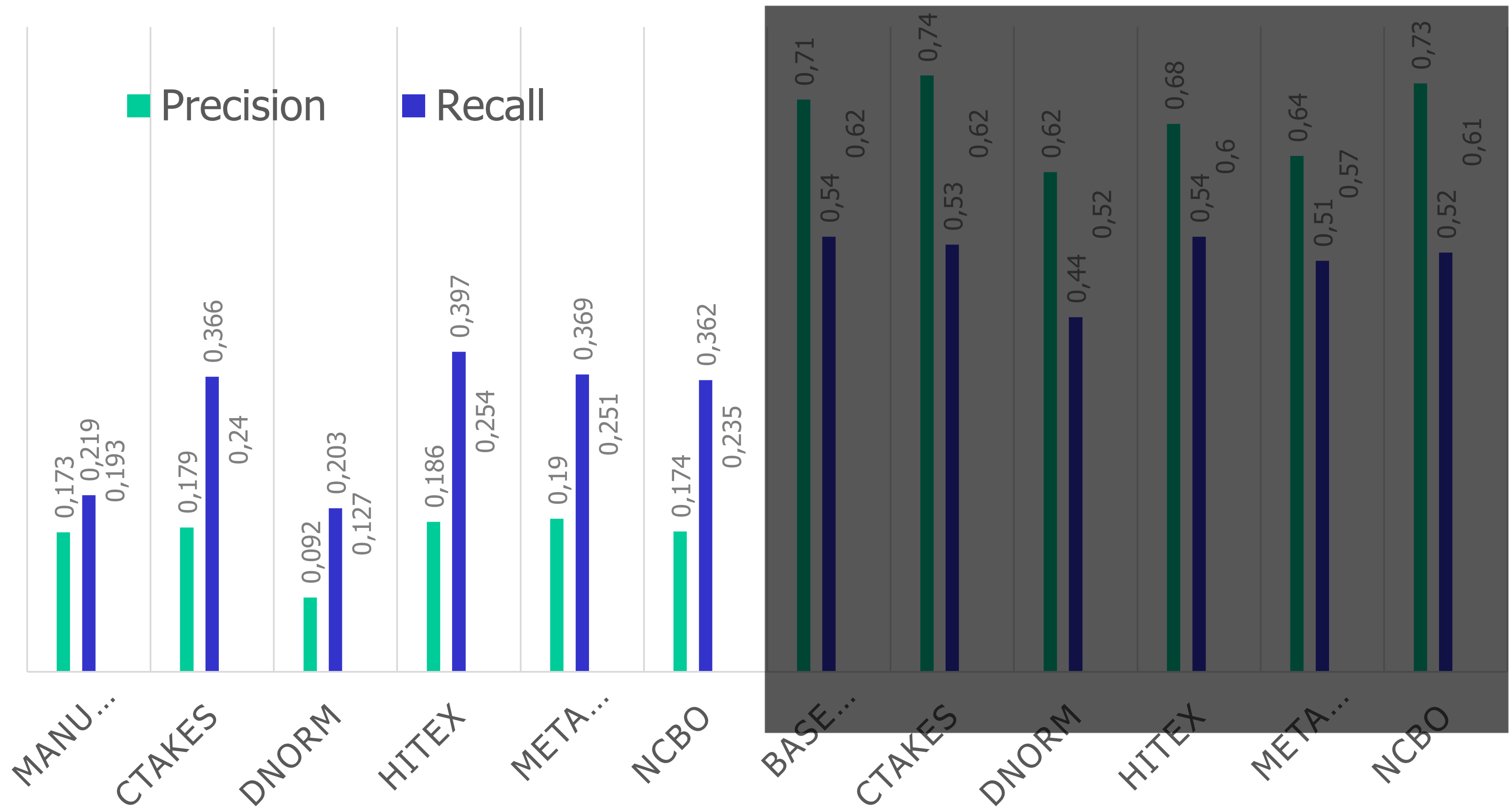
Issues (Typical)

- 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
  - …



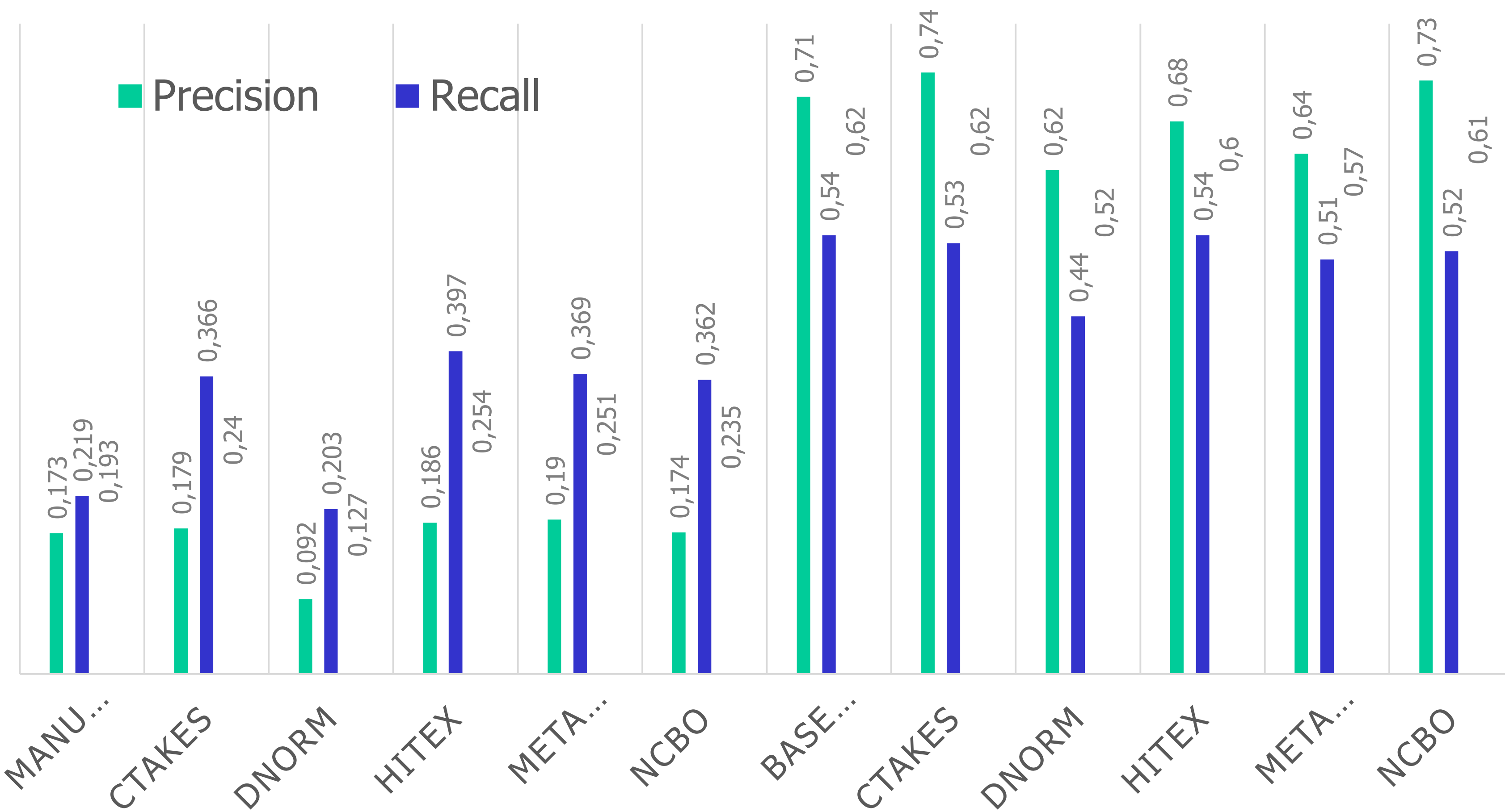
Results / Evaluation

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



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 parsing
  - Pragmatic structure, co-reference resolution, …
- Spoken language
- Machine translation
- Cross-language search / analysis
- User interfaces
- Special classification problems: Sentiment analysis etc.
- Question answering
- Topic modelling
- …