Die Kombination von linguistischen und heuristischen Methoden für die Extraktion von Mehrwort-Termen

Bastian Bolender, Sebastian Kreß, Jannik Strötgen

Studienprojekt an der
Ruprecht-Karls-Universität Heidelberg
Seminar für Computerlinguistik

Agenda

- Was sind Mehrwort-Terme
- Projektbechreibung
- Foundations: Terminology ---raus
- Beispiel: Der MeSH Thesaurus
- Automatisches Indexieren
- Teilprojekte
  - Regelbasierter linguistischer Ansatz (Jannik)
  - Pattern Matching Ansatz (Sebastian)
  - Evaluierung (Bastian)

Was sind Mehrwort-Terme (MWT)

- Mehrere Wörter gehören zu einem Ausdruck
- Keine Probleme, wenn die MWT so in Texten vorkommen, wie sie bekannt sind

- Unsere Aufgabe: diskontinuierliche MWT

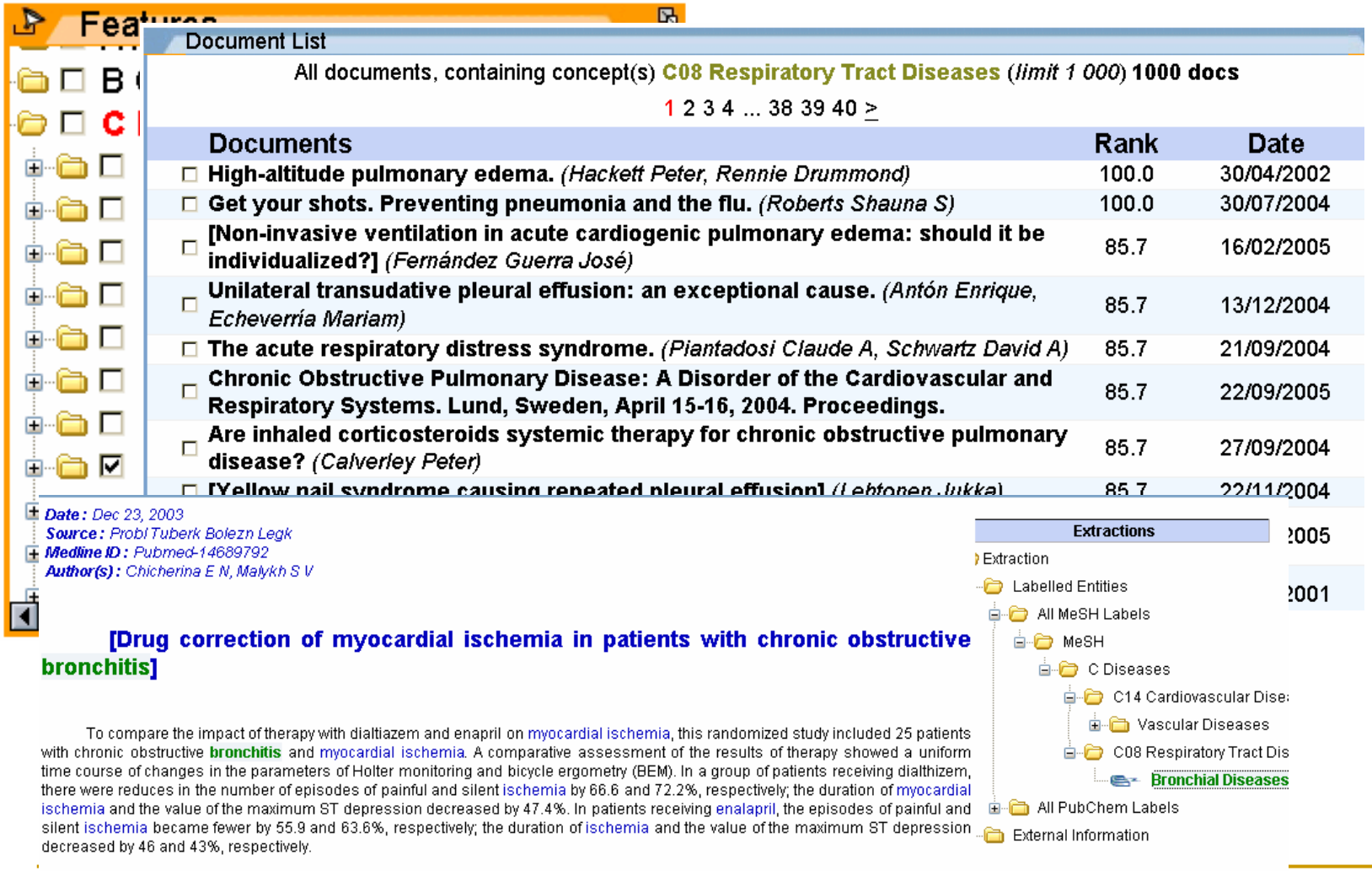
Projektbeschreibung - Einblick

- Automatisches Indexieren von Textdokumenten mit Termen eines kontollierten Vokabulars
- Temis entwickelt Software für automatisches Indexieren
- Frage: Kann die Qualität / Abdeckung durch zusätzliche Methoden verbessert werden
  - Linguistische Methoden
  - Pattern Matching Methoden

Thesaurus: MeSH

- MeSH (Medical Subject Headings): Standard zum Indexieren biomedizinischer Dokumente
- National Institute of Health, Maryland, USA
- Mehrere hunderttausend Terme in komplexer hierarchischer Struktur
- Vgl. http://www.nlm.nih.gov/mesh

MeSH


Document List
All documents, containing concept(s) C08 Respiratory Tract Diseases (limit 1 000) 1000 docs
1 2 3 4 ... 38 39 40 ≥
Documents
Rank
Date
High-altitude pulmonary edema. (Hackett Peter, Rennie Drummond)
100.0
30/04/2002
Get your shots. Preventing pneumonia and the flu. (Roberts Shauna S)
100.0
30/07/2004
[Non-invasive ventilation in acute cardiogenic pulmonary edema: should it be individualized?] (Fernández Guerra José)
85.7
16/02/2005
Unilateral transudative pleural effusion: an exceptional cause. (Antón Enrique, Echeverría Mariam)
85.7
13/12/2004
The acute respiratory distress syndrome. (Piantadosi Claude A, Schwartz David A)
85.7
21/09/2004
Chronic Obstructive Pulmonary Disease: A Disorder of the Cardiovascular and Respiratory Systems. Lund, Sweden, April 15-16, 2004. Proceedings.
85.7
22/09/2005
Are inhaled corticosteroids systemic therapy for chronic obstructive pulmonary disease? (Calverley Peter)
85.7
27/09/2004
Date : Dec 23, 2003
Source : Probl Tuberk Bolezn Legk
Medline ID : Pubmed-14689792
Author(s) : Chicherina E N, Malykh S V
[Drug correction of myocardial ischemia in patients with chronic obstructive bronchitis]
To compare the impact of therapy with dialtiazem and enapril on myocardial ischemia, this randomized study included 25 patients with chronic obstructive bronchitis and myocardial ischemia. A comparative assessment of the results of therapy showed a uniform time course of changes in the parameters of Holter monitoring and bicycle ergometry (BEM). In a group of patients receiving dialthizem, there were reduces in the number of episodes of painful and silent ischemia by 66.6 and 72.2%, respectively; the duration of myocardial ischemia and the value of the maximum ST depression decreased by 47.4%. In patients receiving enalapril, the episodes of painful and silent ischemia became fewer by 55.9 and 63.6%, respectively; the duration of ischemia and the value of the maximum ST depression decreased by 46 and 43%, respectively.
Extractions
Extraction
Labelled Entities
All MeSH Labels
MeSH
C Diseases
C14 Cardiovascular Dise
Vascular Diseases
C08 Respiratory Tract Dis
Bronchial Diseases
All PubChem Labels
External Information

Anwendung kontrollierten Vokabulars

- Ziel: Automatsches Indexieren: Welche Terme sind Deskriptoren fuer ein gegebenes Dokument?
  - Homogenität des Indexierungsprozesses

- Herausforderung
  - Terme werden oft nicht wörtlich in den Texten gefunden.

MeSH Terms

MeSH Term

Diabetes Mellitus, Type 2

Diabetes Mellitus, Adult-Onset
Diabetes Mellitus, Ketosis-Resistant
Diabetes Mellitus, Maturity-Onset
Diabetes Mellitus, Non-Insulin-Dependen
Diabetes Mellitus, Slow-Onset
Diabetes Mellitus, Stable
MODY
Maturity-Onset Diabetes Mellitus
NIDDM
Diabetes Mellitus, Non Insulin Dependent
Diabetes Mellitus, Noninsulin Dependent
Diabetes Mellitus, Type II
Type 2 Diabetes Mellitus

MeSH beinhaltet viele Synonyme

aber was ist mit:

„Type II Diabetes Mellitus“
„Diabetes Mellitus (Slow-Onset)“
....

Synonyme

Erweiterungen I

- Singular / Plural
  - „substance“ / „substances“
- Verschiedene Orthographien
  - „Diabetes Mellitus“ / „diabetes mellitus“
  - „TNF-alpha“ / „TNF alpha“ / „TNF-α“
- Syntaktische Variationen
  - Modifikationen mit Adjektiven, Attributen, usw.

Erweiterungen II

Weitere Erweiterungen sind notwendig:

„... *conducted a structural and functional analysis...*“

→ „structural analysis“

„... *developed different types of cancer. Especially the colon was affected* ....“

→ „colon cancer“

→ Dies ist das Thema unseres Studienprojektes in Zusammenarbeit mit TEMIS / Fraunhofer SCAI

Regelbasierter linguistischer Ansatz

Jannik Strötgen
Ruprecht-Karls-Universität Heidelberg
Seminar für Computerlinguistik

Regelbasierter linguistischer Ansatz

- Grundlagen für regelbasierte linguistische Anwendungen
- Beschreibung der Input-Texte und der MeSH-Datenbank
- Drei Hauptprobleme
  - Permutationen
  - Insertionen (mit Permutationen)
  - Enumerationen (Aufzählungen)
- Beschreibung des Regelapparats
- Beispiele
- Erste manuelle Evaluierung
- Pro und Contra

Grundlagen für regelbasierte linguistische Verfahren

- Part of Speech (POS) Tagger (Xelda®):
  - Jedem Wort wird seine Wortart (POS-Tag) zugewiesen
    (z.B. Nomen, Artikel, …)
- Lemmatizer (Xelda®):
  - Jedes Wort bekommt seine Grundform (Lemma) zugewiesen
    (z.B. played → play; plays → play)

Die Input-Texte

- Die Input-Texte werden getaggt und lemmatisiert

<PubmedArticle>
…
<PMID>15912976</PMID>
…
<AbstractText>A 14-year-old male presented with abdominal pain, diarrhoea and a sensation of something prolapsing through the anus during defecation, …..</AbstractText>
….

en werden hinzugefügt

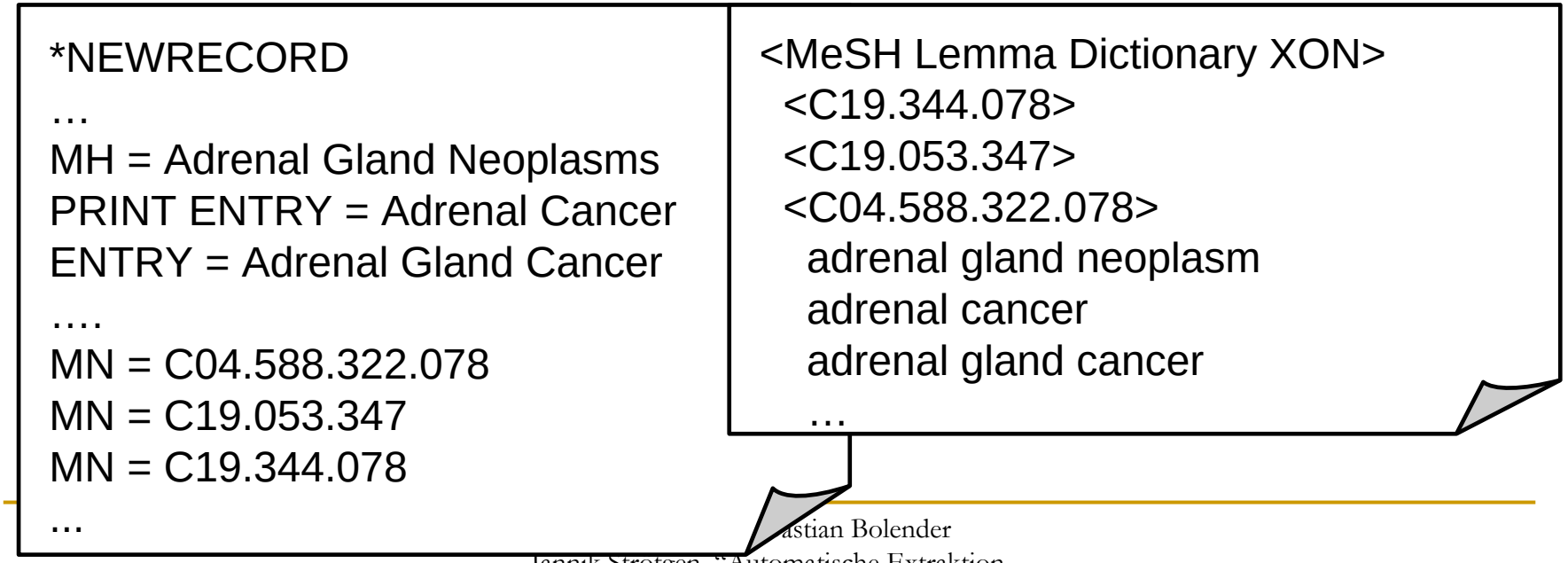
<Xeldaresults file="15912976">
<doc id ="15912976">
<l s="0" pos="#DET">a</l>
<l s="1" pos="#ADJ#GUESSED">14-year-old</l>
<l s="2" pos="#NOUN">male</l>
<l s="3" pos="#VPAP">present</l>
<l s="4" pos="#PREP">with</l>
<l s="5" pos="#ADJ">abdominal</l>
<l s="6" pos="#NOUN">pain</l>
…

Die MeSH-Datenbank

- Mesh-Datenbank:

  Die Terme werden lemmatisiert, damit ein einfacherer Abgleich möglich ist.

  (z.B. Singular / Plural muss nicht berücksichtigt werden)



Drei Hauptprobleme

analogue scale visual
visual analogue scale

- 1. Permutationen *"… analogue visual scale …"*
  - Die Kombinationen müssen errechnet werden
  - Sehr rechenaufwändig,
  - Beschränkung durch maximale Länge

brain injury
brain region injury

- 2. Insertionen (mit Permutationen)

"… injury of brain region …"

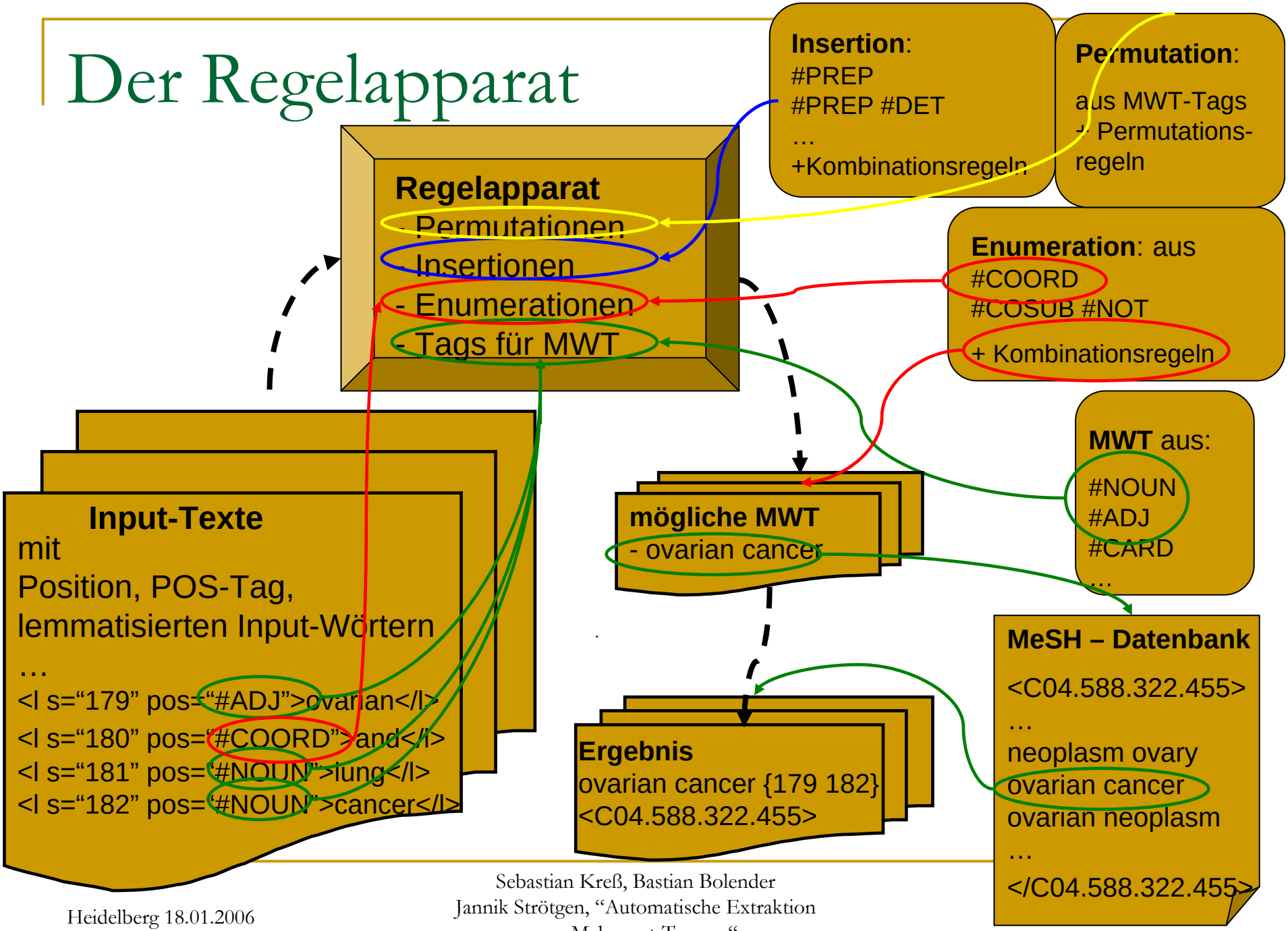
  - mögliche Kombinationen müssen errechnet werden
  - unter anderem Präpositionen verbinden Teile des Mehrwort-Terms (MWT)

Drei Hauptprobleme

structural analysis

- 3. Aufzählungen *"… structural and functional analysis …"*
  - Teil eines Ausdrucks gehört zu mehreren Termen
  - anderer Teil zwischen Anfang und Ende des Ausdrucks muss für neues MWT ignoriert werden

Der Regelapparat



Beispiele – Enumeration

- ... agent in the salvage setting in ovarian, non-small cell lung, breast and colorectal cancers.

(PubMed-ID 16050796)

- ovarian cancer in MeSH (C04.588.322.455)
- breast cancer in MeSH (C04.588.180)
- non-small cell lung cancer
  - cell lung cancer
  - lung cancer

} mögliche Kombinationen

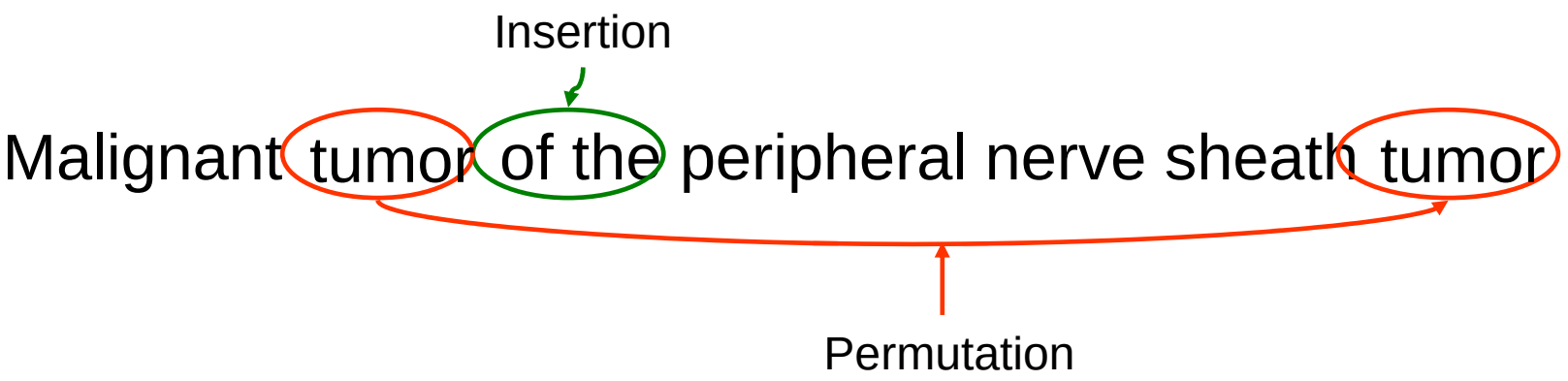
Ausgabe entweder „found MWT“ oder „possible MWT“

“lung cancer” in MeSH (C04.588.894.797.520)

“non-small cell lung cancer”

Beispiele – Insertion mit Permutation

- Malignant tumors of the peripheral nerve sheath are most commonly … (PM-ID 15915634)
  - Malignant peripheral nerve sheath tumor
  - Gefunden in MeSH (C04.557.580.600)



Erste Evaluierung I

- 100 zufällig ausgewählte Abstracts von Pubmed 2004 (ø 214 Wörter und Satzzeichen)

MWT gefunden	wörtlich in Mesh-DB	Synonym aus DB im Text vorhanden	richtig	neuer Index
Enumer.: 20	15	8	20	7
Insertion: 5	2	1	3	1

- Recall: ? (nicht abschätzbar, da keine Annotation)
- Precision: ~ 100% bei exaktem Abgleich mit MeSH
  - Unberücksichtigt: richtige MWT, die nicht in Mesh-DB sind

Erste Evaluierung II

- Enumerationen
  - Relativ feste Regeln
  - gute Ergebnisse allgemein: sehr viele richtige MWTs, die nicht in MeSH-DB stehen
  - mit Abgleich in Mesh weniger Ergebnisse, aber “nur” richtige MWT
- Insertionen / Permutationen
  - einzelne Regeln komplex, nicht so eindeutig wie bei Enumerationen
  - Ohne Mesh-Abgleich viele falsche MWTs, wegen viel mehr Kombinationsmöglichkeiten
  - Regeln und mögliche Kombinationen noch zu verbessern

Pro und Contra

- Contra
  - Nur Konstruktionen, die sich in Regeln ausdrücken lassen, können gefunden werden
  - Taggen und Lemmatisieren kostet Zeit
- Pro
  - Möglichkeit zur linguistischen Weiterverarbeitung
    - gefundene MWT können wegen Positionsangaben und syntaktischen Umformungen weiterverarbeitet werden
  - Wenig false positives, denn die neuen MWT resultieren aus regelbasierten linguistischen Umstellungen und werden mit der Mesh-Datenbank abgeglichen.