Quantitativer Mustererkennungsansatz

Sebastian Kreß

Ruprecht-Karls-Universität Heidelberg

Seminar für Computerlinguistik

Agenda

- Analyse und Zerlegung der *MeSH*-Thesaurus-Daten
- Lattice-Entstehung aus *Pubmed*-Abstracts
- Erweiterte Mustererkennung in verschiedenen Präzisionen und Beispiele
- Verbesserung der Präzision durch
  - Gewichtung von Eingabewörtern mit *tf-idf*
  - Abschätzung mit *semantischer Distanz*
- Probleme und Ausblick

MeSH: Analyse und Zerlegung

chemical	-> MRI(1)
shift	-> MRI(1)
imaging	-> MRI(0,1,3)
mr	-> MRI(2)
tomography	-> MRI(2)
magnetization	-> MRI(3)
transfer	-> MRI(3)
contrast	-> MRI(3)

MeSH-Heading

Magnetic Resonance Imaging

Entry Terms

Chemical Shift Imaging

MR Tomography

Magnetization Transfer Contrast Imaging

[..]

Lattice-Entstehung

Magnetic resonance (MR) imaging revealed an anteromedial temporal..

```
<word nr="50" form="imaging">
   <concept>E07.671/Phantoms, Imaging</concept>
   <concept>E01.370.350.850/Ultrasonography</concept>
   <concept>E01.370.350.500.510/Magnetic Resonance Imaging, Cine</concept>
   <concept>E01.370.350.500.200/Echo-Planar Imaging</concept>
   <concept>E01.370.350.400/Imaging, Three-Dimensional</concept>
   <concept>E01.370.350.710.715.710.350/Gated Blood-Pool Imaging</concept>
   <concept>E01.370.350/Diagnostic Imaging</concept>
   <concept>E01.370.350.515.402.580.500/Microscopy, Energy-Filtering Transmission Electron</concept>
   <concept>E01.370.350.710/Radionuclide Imaging</concept>
   <concept>E01.370.350.500.150/Diffusion Magnetic Resonance Imaging</concept>
   <concept>E01.370.350.500/Magnetic Resonance Imaging</concept>
   <concept>E01.370.350.500.500/Magnetic Resonance Angiography</concept>
</word>
```

```
   <concept>E01.370.350.500/Magnetic Resonance Imaging</concept>
   <concept>E05.196.890/Surface Plasmon Resonance</concept>
   <concept>E01.370.350.500.500/Magnetic Resonance Angiography</concept>
</word>
```

Erweiterte Mustererkennung..

- Quantitative Analyse statt Erlauben von Abweichungen
- Definition eines F
- Entfernung aller M
  geringerem Vor
  der Anzahl der
  Entry Term
- (Entfernung von *MeSH Headings* nach syntaktischen oder semantischen Regeln)

Diabetes Mellitus, Type 1

Diabetes Mellitus, Type 2

- im Satz
-im Abstract-Text
- im Abstract-Text und -Titel

…mit relativ engem Fenster im Satz

[..] Biofilms are a major concern for clinicians in the treatment of infectious disease because of their resistance to a wide range of antibiotics. Arbekacin, an aminoglycoside antibiotic, is the drug of choice for the treatment of **infection** caused by methicillin-resistant **Staphylococcus** aureus (MRSA). However, it has not yet been defined whether arbekacin tends to penetrate into the biofilm structure induced by MRSA infection.[..]

(from: Pubmed 2004, PMID 16163460)

Lattice: 1 2 3 … 11 12 13 **x** 15 16 17 18 **x** => 14 19

- Bacterial Infections and Mycoses
  - Bacterial Infections
    - Gram-Positive Bacterial Infections
      - **Staphylococcus Infections**

…mit weitem Fenster im Satz

[..] The authors sought to determine whether the risk of **congenital heart disease** (CHD) was greater for the children of mothers who lived close to a hazardous waste site (HWS) than for those who lived farther away. All **case**s (n = 1283) of confirmed CHD, and a random sample of 2,292 **control**s, born in Dallas County, Texas, from 1979-1984 were linked with 276 HWSs present during the **study**. The authors ascertained locations of households and determined th

(from Pubmed 2004, PMID 16189989)

Lattice: 1 **x** 3 4 5 6 7 8 9 10 11 12 13 **x** … 30 **x** => 2 14 31

- Investigative Techniques
  - Epidemologic Methods
    - Epidemologic Study Characteristics
      - Epidemologic Studies
        - **Case Control Studies**

…über den ganzen Abstract

[..] Cortical motor organization/reorganization was studied in patients with malformation of cortical development (MCD) by applying two noninvasive motor **mapping** techniques: transcranial magnetic stimulation (TMS) and functional magnetic resonance (fMR) imaging. METHODS: Eight patients [illegible] from congenital [illegible] lesions were [illegible] polymicrogyria [illegible] MCDs involved [illegible] hemiparesis. Transcranial magnetic stimulation was used to search, in both hemispheres, for **brain** regions with corticospinal projections to the paretic hand, and cortical activation during simple repetitive movements of the paretic hand was monitored using fMR imaging. [..]

Lattice: 1 2 3 … 17 18 19 **x** 21 … 87 88 89 **x** …=> 20 90

- Investigative Techniques
  - **Brain Mapping**

(from: Pubmed 2004, PMID 16189989)

Verbesserung der Präzision

- Mit der bisherigen Methode
  - *Recall* für unabstrahierte MeSH-Headings bei großem Fenster fast 100%
  - *Precision* sehr schlecht
- Verbesserungsvorschläge
  - Reduzierung und Konkretisierung der Eingabewörter mit *tf-idf* (versucht, aber problematisch)
  - Reduzierung der identifizierten MeSH-Headings per *semantic distance*

Reduzierung der Eingabewörter

- Durch Festlegung von Wörtern, die dauerhaft ignoriert werden
- Durch tf-idf
  - Idee: wichtigste Terme sind die, die häufig in individuellen Dokumenten auftreten und selten im gesamten Korpus
  - Reduzierung auf Schwellwert
- hier problematisch, weil im Korpus auch häufige Wortteile signifikant sind

$$tf - idf = \frac{n(term)}{n(abstracts)} * \log \frac{n(abstracts)}{n(abstracts\_with\_term)}$$

Semantische Distanz

- *Annahme 1*: die MeSH-Headings in einem Abstract sind semantisch nah (bzw. es gibt Gruppen von semantisch nahen MHs)

- *Annahme 2*: es gibt eindeutig identifizierbare MeSH-Headings (singleword terms oder kontinuierliche multiword terms)

➔ Die Wahrscheinlichkeit für unsichere MeSH-Headings kann aus der semantischen Distanz zu sicheren MeSH-Headings abgeleitet werden

Berechnung (in Taxonomien)

- Lin (1998)

$$\mathrm{sim}_{\mathrm{L}}(c_1, c_2) = \frac{2 \times \log p(lso(c_1, c_2))}{\log p(c_1) + \log p(c_2)}$$

lowest super ordinate

Wahrscheinlichkeit der Elemente im Vergleich zum nächsten

- Leacock and Chodorow (19

$$\mathrm{sim}_{\mathrm{LC}}(c_1, c_2) = -\log \frac{\mathrm{len}(c_1, c_2)}{2D}$$

kürzeste Distanz zwischen c_1 und c_2

Tiefe der Taxonomie

Probleme

- *semantic distance* erfordert Disambiguierung der MeSH Headings zu Konzepten
- Vo
- Fe
  Lä
- Id

Lung Neoplasms

Neoplasms / Neoplasms by Site / Thoracic Neoplasms / Respiratory Tract Neoplasms / Bronchial Neoplasms / Lung Neoplasms [C04.588.894.797.520]

Respiratory Tract Diseases / Lung Diseases / Lung Neoplasms [C08.381.540]

Respiratory Tract Diseases / Respiratory Tract Neoplasms / Lung Neoplasms [C08.785.520]