Large-Scale Evaluation of a Medical Cross-Language Information Retrieval System

Kornél Markó\textsuperscript{1,2}, Philipp Daumke\textsuperscript{1,2}, Stefan Schulz\textsuperscript{2}, Rüdiger Klar\textsuperscript{2}, Udo Hahn\textsuperscript{3}

\textsuperscript{1}Averbis GmbH, Freiburg, Germany
\textsuperscript{2}Medical Informatics Department, University Medical Center Freiburg, Germany
\textsuperscript{3}Jena University Language and Information Engineering Lab (JULIE), Germany
Medical Information

Heterogeneous User Groups

- Physicians, Scientists
- Nurses, Laymen

Heterogeneous Document Types

- Discharge Summaries
- Pathology Reports
- Scientific Publications, etc.

Language Variability

- Experts vs. Laymen
- Linguistic Morphology
- Cross-Lingual

User Interface

Design?
Accessibility?
Bridging the gap between user groups and information sources?
Risk factor high blood pressure

Risk factors hypertension

Risk factor hypertension
User Interface

Korrelation von Hypertonie und Läsion der Weißen Substanz...
Interaction between hypertension, apoE, and cerebral white matter lesions.

d de Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Gijn J, Breteler MM.

Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral white matter lesions (WMLs) are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially hypertension, are related to their presence. However, not every person with vascular risk factors has WMLs, which suggests interaction with other determinants, eg, genetic factors. The epsilon4 allele...
Interaction between hypertension, apoE, and cerebral white matter lesions.

dee Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Gijn J, Breteler MM.

Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral white matter lesions (WMLs) are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially hypertension, are related to their presence. However, not every person with vascular risk factors has WMLs, which suggests interaction with other determinants, eg, genetic factors. The epsilon4 allele
Interaction between hypertension, apoE, and cerebral white matter lesions.

de Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Gijn J, Breurker MM.

Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral white matter lesions (WMLs) are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially hypertension, are related to their presence. However, not every person with vascular risk factors has WMLs, which suggests interaction with other determinants, eg, genetic factors. The epsilon-4 allele
Linguistic phenomena adversely influence medical text retrieval!

- **Inflection**: leukocyte vs. leukocytes, appendix vs. appendices
- **Derivation**: leukocyte, vs. leukocytic
- **Composition**: leuk | em | ia, para | sympath | ectomy
  Magen | schleim | haut | entzünd | ung
- **Acronyms**: AIDS, SARS, OECD
- **Orthographic Variants**:
  - Kolonkarzinom, Coloncarcinom,
  - Esophagus, Oesophagus,
- **Synonyms**:
  - High blood pressure – Hypertension,
  - Prophylaxis – Prevention
Subword-based, multilingual semantic indexer for document retrieval

Subwords are atomic, conceptual or linguistic units:

- Stems: stomach, gastr, diaphys
- Prefixes: anti-, bi-, hyper-
- Suffixes: -ary, -ion, -itis
- Infixes: -o-, -s-

Equivalence classes contain synonymous subwords and their translations:

- #derma = { derm, cutis, skin, haut, kutis, pele, cutis, piel, ... }
- #inflamm = { inflamm, -itic, -itis, -phlog, entzuend, -itis, -itisch, inflam, flog, inflamm, flog, ... }
Morphosaurus Structure

- Thesaurus:
  ~21,000 equivalence classes (MIDs)
- Lexicon entries:
  - English: ~23,000
  - German: ~24,000
  - Portuguese: ~15,000
  - Spanish: ~11,000
  - French: ~8,000
  - Swedish: ~10,000

Segmentation:
Myo | kard | itis
Herz | muskel | entzünd | ung
Inflamm | ation of the heart muscle

Indexation:
#muscle #heart #inflamm
#heart #muscle #inflamm
#inflamm #heart #muscle
<table>
<thead>
<tr>
<th>Original Document</th>
<th>Orthographic Normalization</th>
<th>Morphological Segmentation</th>
<th>Semantic Normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>High TSH values suggest the diagnosis of primary hypothyroidism while a suppressed TSH level suggests hyperthyroidism.</td>
<td>high tsh values suggest the diagnosis of primary hypothyroidism while a suppressed tsh level suggests hyperthyroidism.</td>
<td>high tsh value suggests the diagnosis of primary hypothyroidism while a suppressed tsh level suggests hyperthyroidism.</td>
<td>#up# tsh #value# #suggest# #diagnost# #primar# #small# #thyre# #suppress# tsh #nivell# #suggest# #up# #thyre#.</td>
</tr>
<tr>
<td>Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose, ein supprimierter TSH-Spiegel spricht dagegen für eine Schilddrüsenüberfunktion.</td>
<td>erhoehete tsh-werte erlauben die diagnose einer primaren hypothyreose, ein supprimierter tsh-spiegel spricht dagegen fuer eine schilddruessenueberfunktion.</td>
<td>er heeh te tsh - wert erlaub en die diag nose einer prima ren hypothyreose, ein supprim iert er tsh - spiegel spricht dagegen fuer eine schilddruen eu ber funktion.</td>
<td>#up# tsh #value# #suggest# #diagnost# #primar# #small# #thyre# #suppress# tsh - {#mirror# #nivell#} #say# #thyre# #up# #function#.</td>
</tr>
<tr>
<td>A presença de valores elevados de TSH sugere o diagnóstico de hipotireoidismo primário, enquanto níveis suprimidos de TSH sugerem hipertireoidismo.</td>
<td>a presença de valores elevados de tsh sugere o di gnostico de hipotireoidismo primario, enquanto niveis suprimidos de tsh sugerem hipertireoidismo.</td>
<td>a pr esenca de val ores elevados de tsh sugere o diagnostico de hipotireoidismo primario, enquanto niveis suprimidos de tsh sugerem hipertireoidismo.</td>
<td>#act# #value# #up# tsh #suggest# #diagnost# #small# #thyre# #primar# #nivell# #suppress# tsh #suggest# #up# #thyre#.</td>
</tr>
</tbody>
</table>
Disambiguation

- Maximum likelihood estimator
- Co-occurrence information from large heterogeneous corpora
- `#patient` should be preferred over `#patience`, since „Patient“ is unambiguous in German and also co-occurs with `#heart`

---

The American Heart Association recommends aspirin use for any patient who had a heart attack.

```plaintext
The American Heart Association recommends aspirin use for any patient who had a heart attack.

In Deutschland leben 120.000 erwachsene Patienten mit angeborenen Herzfehlern.
```

```plaintext
#america #heart #associat #advice
#aspirin #utilis { #patient, #patience } #heart #attack.

#german #life 120.000 #adult
#patient #heredit #heart #failure.
```
Interaction between hypertension, apoE, and cerebral white matter lesions.

de Leeuw FE, Richard F, de Groot JC, van Duijn CM, Hofman A, Van Gijn J, Breteler MM.

Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands.

BACKGROUND AND PURPOSE: Cerebral white matter lesions (WMLs) are frequently found on magnetic resonance imaging scans in both cognitively intact and demented elderly persons. Vascular risk factors, especially hypertension, are related to their presence. However, not every person with vascular risk factors has WMLs, which suggests interaction with other determinants, e.g., genetic factors. The epsilon4 allele
Korrelation von Hypertonie und Läsion der Weißen Substanz...

#correl #hyper #tens #lesion #whit #matter

de leeuw fe, richard f, de groot jc, van duijn cm, hofman a, van gijn j, breteler mm.

#department #epidem #logic #bio #statist, erasmus #medic #centr, rotterdam, #dutch.

#rock #ground #purpos : #cerebr #whit #matter #lesion (wmls) #frequent #find #magnet #resonance #imag #scan #both #cognit #dement #gero #human. #vascul #risk #factor, #special #hyper #tens, #relat #presenc. #not #total #human #vascul #risk #factor wmls, #suggest #interact #other #determin, eg,
Korrelation von Hypertonie und Läsion der Weißen Substanz...

#correl #hyper #tens #lesion #whit #matter
Evaluation

- Gold standards: OHSUMED, ImageCLEFMed

- OHSUMED-Corpus (Hersh et al., 1994)
  - Subset of MEDLINE
  - ~233,000 English documents
  - 106 English user queries

- ImageCLEFMed Corpus (Clough et al., 2005)
  - Multilingual Image Retrieval Task 2006
  - ~41,000 Medical Images and captions
  - 30 queries

- Query-document pairs have been manually judged for relevance

- Non-English queries obtained by translation to German, Portuguese, Spanish and Swedish by domain experts

- Search Engine: Lucene
  - http://lucene.apache.org/
Evaluation

• **Baseline:** monolingual text retrieval
  – (stemmed) English user queries
  – (stemmed) English texts

• **Query translation (QTR)**
  – Google translator
  – Multilingual dictionary compiled from UMLS

• **Morphosaurus Indexing (MSI)**
  – Interlingual representation of both user queries and documents
  – **MSI-D** incorporating disambiguation module
Results: Ohsumed

Mean Average Precision

Percent of Baseline

Language and Condition

Query Translation
Morphosaurus
Morphosaurus+D
Results: ImageCLEFMed

Top 20 Average Precision

Percent of Baseline

Language and Condition

EN DE PT SP FR SV AV

Query Translation
Morphosaurus
Morphosaurus+D
Conclusions

• Cross-Language Document Retrieval
  – Based on morphological and semantic normalization of both user queries and documents
  – Matching of search/document terms on a language-independent, interlingual layer

• Language-independent indexing
  – reaches more than 92% of an English-English baseline on heterogeneous document collections, in average
  – outperforms query translation significantly
  – is independent from particular search engine architectures

• Morphosaurus incorporates six languages:
  – German, English, Portuguese, Spanish, French, Swedish

• In use in commercial systems
Contact

Stefan Schulz
Medical Informatics Department
University Hospital Center Freiburg, Germany
stschulz@uni-freiburg.de

www.imbi.uni-freiburg.de

Kornél Markó
Averbis GmbH, Freiburg, Germany
marko@averbis.de

www.averbis.net