Presenter Name: Stefan Schulz
Country: 1. Germany, 2. Brazil
Qualification(s):
MD (Doctor in Theoretical Medicine)
Vocational Training in Medical Informatics
Postdoctoral Habilitation degree in Medical Informatics
Position:
Associate Professor
Department/ Organisation:
1. Medical Informatics, Freiburg University Medical Center, Freiburg, Germany
2. Master Program of Health Technology, Catholic University of Paraná, Curitiba, Brazil
Major Achievement(s):
Research in the fields of Medical Terminology, Biomedical Ontologies, Medical Language Processing, Text Mining, and Document Retrieval.
Medical Thesaurus Anomaly Detection by User Action Monitoring

Jeferson L. Bitencourt\textsuperscript{a,b}, Pindaro S. Cancian\textsuperscript{a,b},
Edson Pacheco\textsuperscript{a,b}, Percy Nohama\textsuperscript{a,b}, Stefan Schulz\textsuperscript{b,c}

\textsuperscript{a} Paraná University of Technology (UTFPR), Curitiba, Brazil
\textsuperscript{b} Pontifical Catholic University of Paraná, Master Program of Health Technology, Curitiba, Brazil
\textsuperscript{c} University Medical Center Freiburg, Medical Informatics, Freiburg, Germany
Thesaurus

- Controlled Vocabulary for document indexing and retrieval
- Assigns semantic descriptors (concepts) to (quasi-)synonymous terms
- Contains additional semantic relations (e.g. hyperonym / hyponym)
- Examples: MeSH, UMLS, WordNet
- Multilingual thesaurus: contains translations (cross-language synonymy links)
Multilingual Thesaurus Management

- International team of lexicon curators
- React to new terms and senses
- Decide which terms are synonymous / translations
- Decide which senses of a term have to be accounted for in the domain
- Requires quality assurance measures
Case study: Morphosaurus

- Medical subword thesaurus
- Organizes subwords (meaningful word fragments) in multilingual equivalence classes:
  - #derma = \{ derm, cutis, skin, haut, kutis, pele, cutis, piel, \ldots \}
  - #inflamm = \{ inflamm, -itic, -itis, phlog, entzuend, -itis, -itisch, inflam, flog, inflam, flog, \ldots \}
- Maintained at two locations: Freiburg (Germany), Curitiba (Brazil)
- Lexicon curators: frequently changing team of medical students
Morphosaurus Structure

- Thesaurus: ~21,000 equivalence classes
- Lexicon entries:
  - English: ~23,000
  - German: ~24,000
  - Portuguese: ~15,000
  - Spanish: ~11,000
  - French: ~8,000
  - Swedish: ~10,000
<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 s suggest the diagnosis is of primar y hypo thyroid ism while a suppress ed tsh level suggest s hyper thyroid ism.</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 primaer hypothyreose, ein supprimierter tsh-spiegel spricht dagegen fuer eine schilddruenessuberfunktion.</td>
<td>er hoeh te tsh - wert e erlaub en die diagnos e einer primaer en hypo thyre ose, ein supprim iert er tsh - spiegel spricht dagegen fuer eine schilddruesenuberfunktion.</td>
<td>#up# tsh - #value# #permit# #diagnost# #primar# #small# #thyre# , #suppress# tsh - {#mirror# #nivell#} #speak# #thyre# #up# #function# .</td>
</tr>
<tr>
<td>A presencia de valores elevados de TSH sugere o diagnóstico de hipotireoidismo primário, enquanto níveis suprimidos de TSH sugerem hipertireoidismo.</td>
<td>a presencia de valores elevados de tsh sugere o diagnostico de hipotireoidismo primario, enquanto niveis suprimidos de tsh sugerem hipertireoidismo.</td>
<td>a presene a de valor es elevad os de tsh suger e o diagnost i co de hipo tireoid ismo pri mari o, enquanto niveis suprimid os de tsh suger em hiper tireoid ismo.</td>
<td>#actual# #value# #up# tsh #suggest# #diagnost# #small# #thyre# #primar# , #nivell# #suppress# tsh #suggest# #up# #thyre# .</td>
</tr>
</tbody>
</table>
Morphosaurus: 2 Semantic Relations

Composition: *Has_word_part*

Specialization: *Has_sense*
<table>
<thead>
<tr>
<th>Introduction</th>
<th>Methods</th>
<th>Results</th>
<th>Discussion</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Morphosaurus Building Pragmatics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Properly delimit subword entries so that they are correctly extracted from complex words:

- `nephrotomy` -> `nephro` | `oto` | `my`
- `nephrotomy` -> `nephro` | `tomy`
- Properly delimit subword entries so that they are correctly extracted from complex words:
  - nephrotomy -> nephr / oto / my
  - nephro	omy -> nephro / tomy

- Create consensus about the scope of synonymy classes, especially with regard to highly ambiguous words
Morphosaurus Quality Assurance

- Content quality: Identify content errors in the thesaurus content (see Andrade et al., MEDINFO 2007)
- Process quality: Detect and prevent user action anomalies
- User action anomalies: actions that consume effort without any positive impact: uncoordinated edit / update / delete “do undo” transactions done by different lexicographers
Identification of Editing Anomalies

• Analysis of data logs patterns: 86 thesaurus backups covering 9 months
• Assessing relevance of anomaly patterns by comparing the thesaurus descriptors affected with those debated in a Morphosaurus editor online forum
Identification of Editing Anomalies

• Analysis of data logs patterns: 86 thesaurus backups covering 9 months

• Assessing relevance of anomaly patterns by comparing the thesaurus descriptors affected with those debated in a Morphosaurus editor online forum
Anomalies: Typology

1. Relationship anomaly
2. Type Anomaly
Anomalies: Typology

3. Delimitation Anomaly
4. Permanence anomaly
Identification of Editing Anomalies

- Analysis of data logs patterns: 86 thesaurus backups covering 9 months
- Assessing relevance of anomaly patterns by comparing the thesaurus descriptors affected with those debated in a Morphosaurus editor online forum
Example of Morphosaurus forum entry

EqClass spotted by corpus based content quality analysis, cf. Andrade et al., MEDINFO 2007
<table>
<thead>
<tr>
<th>Anomaly Type</th>
<th>Occurrences</th>
<th>Discussed in Forum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relationship anomaly</td>
<td>76</td>
<td>28</td>
</tr>
<tr>
<td>Type anomaly</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Delimitation anomaly</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Permanence anomaly</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
### Relationship anomalies: multiple changes

<table>
<thead>
<tr>
<th>Number of do-undo actions</th>
<th>Occurrences</th>
<th>Discussed in Forum</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>23</td>
<td>10</td>
</tr>
</tbody>
</table>
Problems found by Log Analysis

<table>
<thead>
<tr>
<th>Problem Type</th>
<th>Occurrences</th>
<th>Found by Log Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>An expected relation relating ambiguous or expansible semantic identifiers (\textit{has_sense} type or \textit{has_word_part} type)</td>
<td>86</td>
<td>24</td>
</tr>
<tr>
<td>Entries assigned to one semantic identifier did not cover all languages.</td>
<td>80</td>
<td>6</td>
</tr>
<tr>
<td>The same sense is represented by two unrelated semantic identifiers.</td>
<td>70</td>
<td>8</td>
</tr>
<tr>
<td>Lexicon entries assigned to one semantic identifier diverge in meaning.</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Language specific entry do not translate to other languages.</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>Orthographic errors.</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>Similar senses are represented by two unrelated semantic identifiers, one of them of the type “excluded from indexing”.</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>Errors caused by incorrect subword delimitation</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>Errors caused by incorrect functioning of the segmentation engine.</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Discussion of Results

- Assignment of semantic relations: main cause of do-undo anomalies (up to seven do-undos)
- Nearly half of editing anomalies concern semantic identifiers also identified as problematic by corpus analysis
- Problems discussed in forum exceeds those identifiable by log analysis
- Surprising: no anomaly of string delimitation found
Anomaly detection

- Detects waste of resources by “do - undo” actions in thesaurus management
- Helps create consensus in borderline decisions
- Useful to discover common anomalies
- To be complemented by other techniques
- Higher process effectiveness by integration of quality assessment routines in the thesaurus management tools: User alert at runtime
Anomaly detection at runtime

You are undoing a change performed by user koppe on May, 14. Please contact this user and create consensus or discuss the problem at the MorphoSaurus forum!