Evaluating monolingual term extraction from German texts
Tanja Georger, Johannes Schäferer, Ina Rösigier, Ulrich Heider, Michael Dornaer
*Universität Stuttgart, *Universität Hildesheim, ©Robert Bosch GmbH

Framework and objectives
- Experiments on high quality term extraction: Research collaboration, University ↔ BOSCH corporate research
- Domain: German do-it-yourself instructions – DIY expert texts and user-generated content (UGC)
- Tool evaluation:
  - Hybrid research prototype
  - Alternative components:
    - statistical and syntax-based
    - Statistical tool: commercial product (SDL)

Evaluation methodology
- Use of manually designed gold standard:
  - 3 independent annotators: +/- domain specific
- Patterns:
  - N: Bohrmaschine, Schraubenzieher, Loch
  - Adj+N: ossilierende Säge, geborhotes Loch
  - N+N: Führung der Säge, Kopf einer Schraube
  - N+von+N: Fräsen von Kanten, Scheiben von Holz
  - N+Pr+N: Handkreissäge mit Führungsschiene, Spiralbohrer für Metall
- Strict vs. liberal gold standard:
  - Full agreement (3:0) vs. majority vote (2:1)
  - Automatic evaluation: precision, recall, f-measure
- All results collected in a database

DIY corpus
Size and composition of the corpus:
<table>
<thead>
<tr>
<th>text type</th>
<th># of tokens</th>
<th>authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIY manual</td>
<td>62,111</td>
<td>Ulrich</td>
</tr>
<tr>
<td>DIY encyclopedia</td>
<td>8,686</td>
<td>100</td>
</tr>
<tr>
<td>DIY practical Adv</td>
<td>10,104</td>
<td>100</td>
</tr>
<tr>
<td>Marketing tools</td>
<td>30,302</td>
<td>100</td>
</tr>
<tr>
<td>DIY project descrip.</td>
<td>2,100.000</td>
<td>100</td>
</tr>
<tr>
<td>MUD (tumble)</td>
<td>1,150</td>
<td>100</td>
</tr>
<tr>
<td>total</td>
<td>2,728,944</td>
<td></td>
</tr>
</tbody>
</table>

Gold standard development
- Guidelines for cases of doubt: term vs. non-term
  - In-domain vs. out-of-domain ambiguities: Engländler, Rahmen, Leitung, Tor...
  - Abbreviations: PVC, EU
  - Measure indications: 2mm-Bohrer, 240er Schleifpapier vs. 2. Gang, 1-2-do
  - Product and company names: IWO von Bosch
- Size:
  - Pattern: 2,728,944
  - Pos. N: 234,319
  - N+N: 52,387
  - N+von+N: 42,244
  - N+Pr+N: 190,361
  - N+Pr+N: 30,302
  - total: 2,728,944

Evaluation 1: Hybrid vs. statistics only
- Tests:
  - Precision and recall vs. length of candidate list
  - Precision and recall vs. SD quality levels
  - F-measure by SD quality levels
- Evaluation of the POS-pattern N:
  - precision & recall values:
  - IMS precision & recall values:
  - F-measure by IMS prototype:
  - number of extracted term candidates in %
  - number of extracted term candidates in %
  - f-measure values:
  - IMS f-measure

Evaluation 2: Alternative statistical measures
- Tests:
  - Precision, recall, f-measure
  - for all patterns and termhood measures
  - Experiments on combinations of measures

Evaluation 3: Parsing-based extraction
- Motivation:
  - noise in multi-word candidate sets
  - POS patterns: no information about phrase boundaries
  - Example: NP, +Prp+N, should only extract NPs, when NP is embedded in NP: Man legt die Oberfahne nach Arbeitsende ab...
- Method: find start and end points of complex NPs – candidates going beyond phrase boundaries are not counted as valid term candidates
- More extensive gold standard under construction
- Tests and results:
  - Percentage of phrase boundary violations: ca. 8% of token instances of all candidates
  - Manual plausibility check: 83% of top-100 non-term candidates are correctly spotted and removed from result: Vorlage mit Sprühkleber, Schraube zum Einsatz

Conclusions and future work
- So far:
  - Gold-standard-based evaluation: method and database infrastructure
  - First results:
    - Hybrid tool
    - outperforms merely statistical one on DE data
    - MWT noise (f<0.5) can be reduced
      - by use of C-value to ‘correct’ frequency counts
      - - Qualitative results suggest usefulness of phrase boundary filter for MWT extraction
- Future work:
  - More detailed analysis of measure combinations to reduce MWT noise, e.g. termhood plus association measures
- Use of parsing-based extraction:
  - Detailed evaluation of phrase-boundary filter
  - Problem: mate not optimized for phrase boundaries: Experiments also with other parsers
  - Extraction of noun+verb data to find evidence for relational knowledge, e.g. ‘X causes Y’, ‘X uses Y for Z’...
  - Gold standard data in preparation