Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for specialized e-dictionaries

Ina Rösiger, Johannes Schäfer, Tanja George, Simon Tannert, Ulrich Heid, Michael Dorna

Universität Stuttgart, Universität Hildesheim, Robert Bosch GmbH

elex-2015: Herstmonceux, August 2015
Overview

- Context and objectives
- Data and technology:
  Corpus linguistic tools: components and evaluation
- Extraction of relational data from texts:
  taxonomic and non-taxonomic relations between domain objects
- Sample results
- Collecting data for lexicographic purposes
- Conclusion and future work
Objectives

General context

- **Data on the internet:**
  - Domain-specific user-generated content: forums, discussion groups, etc., from the field of do-it-yourself instructions.
  - Expert-produced texts from the same domain: manuals, handbooks, articles, ...

- **Need for professional text analysis:**
  - Tools to analyze the UGC from a domain-related viewpoint: classification by topics, finding answers for (e.g. forum) questions, etc.
  - **Lexical resources to feed the tools:**
    * To be created interactively
    * To be used both interactively and/or automatically
Lexicographic objectives
Identifying raw material for interactive e-dictionary building

- Scenario:
  - Automatic extraction of candidate data from corpora to create entries of a specialized dictionary
    * term candidates
    * term variants – phraseological variants
    * taxonomic and non-taxonomic relations, e.g. “made-of”, “serves-for”...
  - Collecting data for interactive entry construction

![Diagram of corpus preprocessing and term relation extraction]

Results:
- Data for lexicography
- Corpus annotated
- Term relation extraction
- Preprocessing
- Corpus

T&O (IMS/UHI/BOSCH) Terminology tools elex2015-fol.tex 4 / 28
Lexicographic objectives

Focus in this presentation

- Not on dictionary as an end product
- But on tools for
  - term candidate extraction
  - extraction of relational data
- Why not use tools like the SketchEngine?  
  - Relation extraction requires specific procedures
  - Specialized corpora are small: issue for statistical tools
  - Requirements of work on German data:
    * Dependency parsing
    * Analysis of German compounds

Kilgarriff et al. 2004

2.7 - 17.9 M
Case study: Data used
Texts from the do-it-yourself domain (DIY)

- So far:
  opportunistic collection
- Different genres, text types, etc.
- EXP: UGC = 1:4.3
- Subset used in evaluation: 2.7 M

<table>
<thead>
<tr>
<th>Text types</th>
<th>aut.</th>
<th>size (w)</th>
<th>totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIY manuals, tool manuals</td>
<td>EXP</td>
<td>131,254</td>
<td></td>
</tr>
<tr>
<td>DIY (web) encyclopedias</td>
<td>EXP</td>
<td>28,430</td>
<td></td>
</tr>
<tr>
<td>Tool test reports</td>
<td>EXP</td>
<td>239,238</td>
<td></td>
</tr>
<tr>
<td>Marketing texts</td>
<td>EXP</td>
<td>35,302</td>
<td></td>
</tr>
<tr>
<td>DIY articles, “tricks”, etc.</td>
<td>EXP</td>
<td>2,807,487</td>
<td></td>
</tr>
<tr>
<td><strong>Total: expert texts</strong></td>
<td></td>
<td></td>
<td>3,241,711</td>
</tr>
<tr>
<td>DIY project descriptions</td>
<td>UGC</td>
<td>4,479,437</td>
<td></td>
</tr>
<tr>
<td>DIY forum posts</td>
<td>UGC</td>
<td>7,873,115</td>
<td></td>
</tr>
<tr>
<td>Forum FAQs, articles, etc.</td>
<td>UGC</td>
<td>450,143</td>
<td></td>
</tr>
<tr>
<td>Wiki content</td>
<td>UGC</td>
<td>896,267</td>
<td></td>
</tr>
<tr>
<td><strong>Total: user-generated texts</strong></td>
<td></td>
<td></td>
<td>13,698,962</td>
</tr>
<tr>
<td>varia (without metadata)</td>
<td>?</td>
<td>961,236</td>
<td></td>
</tr>
<tr>
<td><strong>total: data collection</strong></td>
<td></td>
<td></td>
<td>17,901,909</td>
</tr>
</tbody>
</table>
Technology used

Term candidate extraction – overview of preprocessing steps

- Standard corpus technology for preprocessing
  - Tokenizing [Schmid 2000]
  - Tagging, Lemmatization: RF-Tagger [Schmid/Laws 2008]
  - Dependency parsing: mate [Bohnet 2010, Björkelund et al. 2010]
Technology used

Term candidate extraction – preprocessing: parsed data

Der Lithium-Ionen-Akku ermöglicht einen von der Steckdose unabhängigen Betrieb des Elektrowerkzeugs

“The Lithium ion accumulator enables an operation of the power tool which is independent from the socket”
Technology used

Term candidate extraction – preprocessing: parsed data

- **Standard dependency representation:**
  - **verb**
  - **subject**
  - **object**
Technology used

Term candidate extraction – preprocessing: parsed data

- Additional tool: extraction from different levels of annotation
  - **heads** of subjects and complements
  - **embedded elements** of subjects and complements
  - **adjuncts** – not part of subjects or complements
Technology used

Term candidate extraction – preprocessing: parsed data

- Extraction from different levels of annotation:
  - **heads** of subjects and complements
  - **embedded elements** of subjects and complements
Technology used

Term candidate extraction – preprocessing: parsed data

• Extraction from different levels of annotation:
  – **heads** of subjects and complements
  – **embedded elements** of subjects and complements
Technology used

Term candidate extraction – preprocessing: parsed data

- Extraction from different levels of annotation:
  - heads vs. embedded elements of subjects and complements

  Information about both:
  Grammatical function and span of sentence constituents

  combined advantages of dependency and constituency
Technology used
Term candidate extraction – patterns and simple statistics

- General overview

  corpus → preprocessing → annotated corpus → term extraction → results: data for lexicography

- Term extraction procedures – two steps:
  extraction by patterns – ranking of extracted candidates

  corpus → preprocessing → pattern search → ranking → term candidate list
Technology used

Term candidate extraction via patterns

- Pattern-based search:
  1. POS-shapes:
     - N Bohrmaschine “drill”
     - Adj + N oszillierende Säge “oscillating drill”
  2. Term-relevant structures extracted from dependency parses:
     - N + PP Bohrer mit Kabel “drill with cord”
     - V + NP
       Temperatur + erhöhen “increase + temperature”

- Relating simple patterns with more complex patterns to find term variants and their relationship with basic terms:
  3. Subtype-denoting: e.g. ((Adv)? (Adj)? Adj)? N:
     Farbe → weisse Farbe “colour → white colour”
  4. Patterns finding cases of embedded term use:
     (N Det)? ((Adv)? Adj)? ((Adv)? Adj)? N
     bodengleiche Dusche → Aufbau einer bodengleichen Dusche
     “walk-in shower → installation of a walk-in shower”
Technological used
Statistical term candidate ranking

• Ranking according to statistical measures: comparison between general-language and domain-specific candidate frequencies
• Domain corpus: DIY data
• General-language corpus: SDeWaC (880 M. tokens)
• Test of several termhood measures
• In current experiments: domain specificity

Schäfer et al. submitted
Ahmad et al. 1992
Output of term extraction: evaluation

Gold standard-based evaluation

• Gold standard (gs)
  – 2.7 M sample from the DIY corpus
  – 3 independent annotators
  – basic patterns only: N, Adj+N, N+N_{Gen}, N+Prp+N
  – Decision: [+/- terminologically relevant]
  – We keep track of \{3:0\}-decisions (strict)
    and of \{2:1\}-decisions (liberal)

• Evaluation experiments:
  – Our tool (basic version) ↔ SDL (Trados) Multiterm Extract
  – Different termhood measures
  – Use of additional (dependency-) syntactic filters

George 2014
Schäfer 2015
Schäfer et al. submitted
Output of term extraction: examples from the evaluation

Quantitative results – comparison with SDL MultiTerm Extract

- Best f-measures per tool

<table>
<thead>
<tr>
<th>pattern:</th>
<th>IMS</th>
<th>SDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>72%</td>
<td>66%</td>
</tr>
<tr>
<td>Recall</td>
<td>84%</td>
<td>68%</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.78</td>
<td>0.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pattern:</th>
<th>IMS</th>
<th>SDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>65%</td>
<td>40%</td>
</tr>
<tr>
<td>Recall</td>
<td>91%</td>
<td>76%</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.76</td>
<td>0.52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pattern:</th>
<th>IMS</th>
<th>SDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>52%</td>
<td>39%</td>
</tr>
<tr>
<td>Recall</td>
<td>85%</td>
<td>76%</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.65</td>
<td>0.52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pattern:</th>
<th>IMS</th>
<th>SDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>38%</td>
<td>33%</td>
</tr>
<tr>
<td>Recall</td>
<td>55%</td>
<td>22%</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.45</td>
<td>0.26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pattern:</th>
<th>IMS</th>
<th>SDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>55%</td>
<td>44%</td>
</tr>
<tr>
<td>Recall</td>
<td>73%</td>
<td>73%</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.63</td>
<td>0.55</td>
</tr>
</tbody>
</table>

- F-measure in terms of quality levels of SDL’s tool:

Noun + Noun_{Genitive}
Extracting data on relations

Overview

• **Taxonomic relations:**
  - Building partial hierarchies of superordinate and subordinate domain objects: by means of taxonomy patterns and compound analysis
  - Including term variants

• **Non-taxonomic relations:**
  Collecting data by means of an analysis of compounds and their syntactic variants
Extracting data on taxonomic relations

Combining different methods

• Taxonomy patterns:
  – *an X is a Y which...*
  – *X₁, X₂,... and other Ys...*
  ⇒ relevant for relations between items that are not morphologically related

• Analysing German compounds:
  X·Y is a type of Y:
  – *Band·säge → Säge*
    band·saw → saw
  – *Mehrzweck·werkbank → Werkbank*
    multi-purpose·workbench → workbench

• Syntactic analysis of phrases expressing taxonomic relations:
  – *Adj+N is a type of N:*
    *durchsichtige Farbe → Farbe*  
    transparent colour → colour
  – *Adv + Adj + N is a type of Adj+N:*
    *matt weiße Farbe → weiße Farbe*  
    matt white colour → white colour
Extracting data on taxonomic relations

Analysis of German compounds – methodology

• Compound splitting with COMPOST, a hybrid tool based on morphological rules and corpus data
  – Head as superordinate
  – Compounds considered as subtypes of their heads:
    \[ \text{Säge} \rightarrow \{ \text{Kreissäge, Bandsäge, ...} \} \]
    \[ \text{saw} \rightarrow \text{circular saw, bandsaw ...} \]

• Implementation is aware of complex non-heads:
  
  (1) Split into morphemes:
    \[ \text{Eigenbaubandsäge} \rightarrow \text{eigen} \cdot \text{bau} \cdot \text{band} \cdot \text{säge} \]
    \[ \text{self-made} \cdot \text{bandsaw} \rightarrow \text{self} \cdot \text{build} \cdot \text{band} \cdot \text{saw} \]

  (2) Check for attested morpheme combinations:
    * Bandsäge
    * *Baubandsäge
    * Eigenbau-X: \text{Eigenbaumöbel, Eigenbauschlitten, etc.}
      
  (3) Correct split: \text{Eigenbau-Bandsäge}
Extracting data on taxonomic relations
Analysis of German compounds – sample results

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandsäge</td>
<td>bandsaw Band</td>
</tr>
<tr>
<td>Elektro-Bandsäge</td>
<td>electric bandsaw Elektro</td>
</tr>
<tr>
<td>Hand-Bandsäge</td>
<td>hand bandsaw Hand</td>
</tr>
<tr>
<td>Horizontalbandsäge</td>
<td>horizontal bandsaw Horizontal</td>
</tr>
<tr>
<td>Vertikalbandsäge</td>
<td>vertical bandsaw Vertikal</td>
</tr>
<tr>
<td>Metallbandsäge</td>
<td>metal bandsaw Metall</td>
</tr>
<tr>
<td>Minibandsäge</td>
<td>mini bandsaw Mini</td>
</tr>
</tbody>
</table>
Extracting data on taxonomic relations
Sample results: Combining patterns and compound analysis

<table>
<thead>
<tr>
<th>← Hypernyms - Hyponyms →</th>
<th>Tools used</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elektrowerkzeug</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Schleifer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Bandschleifer</td>
<td>taxonomic pattern</td>
<td>power tool</td>
</tr>
<tr>
<td>- Exzenterschleifer</td>
<td>compound analysis</td>
<td>sander</td>
</tr>
<tr>
<td></td>
<td></td>
<td>belt sander</td>
</tr>
<tr>
<td></td>
<td></td>
<td>random orbital sander</td>
</tr>
<tr>
<td>Elektrowerkzeug</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Kreissäge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Handkreissäge</td>
<td>taxonomic pattern</td>
<td>power tool</td>
</tr>
<tr>
<td>- Tischkreissäge</td>
<td>compound analysis</td>
<td>circular saw</td>
</tr>
<tr>
<td></td>
<td></td>
<td>circular handsaw</td>
</tr>
<tr>
<td></td>
<td></td>
<td>circular table saw</td>
</tr>
</tbody>
</table>
Extracting data on non-taxonomic relations
Combining compound splitting and the search for syntactic paraphrases

- Compound splitting using COMPOST

- Use of head and non-head items in pattern search:
  different syntactic patterns, depending on type of the head
  - nominal heads:
    - \[ N_1 \cdot N_2 \rightarrow N_2 + \text{Prep} + N_1 \]:
      - \textit{Schraubenloch} \rightarrow \textit{Loch für Schraube} \quad \textit{‘screw·hole’} \rightarrow \textit{hole for screw}
    - \[ N_1 \cdot N_2 \rightarrow N_2 + N_1 - \text{Genitive} \]:
      - \textit{Raummitte} \rightarrow \textit{Mitte des Raums} \quad \textit{‘room·centre’} \rightarrow \textit{centre of room}
  - deverbal heads:
    - \[ N_1 \cdot V_2^n \rightarrow V_2^n + N_1 - \text{Genitive} \]:
      - \textit{Temperaturerhöhung} \rightarrow \textit{Erhöhung der Temperatur} \quad \textit{‘temperature·increase’} \rightarrow \textit{increase of temperature}
    - \[ N_1 \cdot V_2^n \rightarrow V_2 + \text{Obj}(N_1) \]:
      - \textit{Holzbohrer} \rightarrow \textit{Holz + bohren [jmd. bohrt Holz]} \quad \textit{‘wood·drill’} \rightarrow \textit{(to) drill + wood [sbdy drills wood]}
Extracting data on non-taxonomic relations

Purpose and sample results

(1) Getting more evidence for a term candidate:

<table>
<thead>
<tr>
<th>Term Candidate (Component + Component)</th>
<th>Equivalent Term (Component + Component)</th>
<th>$f_{cmpd}$</th>
<th>$f_{synt}$</th>
<th>$\sum$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schraubenloch (screw + hole)</td>
<td>Loch für Schraube (hole for screw)</td>
<td>441</td>
<td>15</td>
<td>456</td>
</tr>
<tr>
<td>Raummitte (room + center)</td>
<td>Mitte des Raumes (center of the room)</td>
<td>37</td>
<td>57</td>
<td>94</td>
</tr>
<tr>
<td>Holzmaserung (wood + grain)</td>
<td>Maserung des Holzes (grain of the wood)</td>
<td>136</td>
<td>56</td>
<td>192</td>
</tr>
<tr>
<td>Brettkante (board + edge)</td>
<td>Kante des Brettes (edge of the board)</td>
<td>79</td>
<td>41</td>
<td>120</td>
</tr>
</tbody>
</table>

(2) Data for specific types of relations:

- **Material**: preposition: *aus* (made of)
  - Stahlschraube | Schraube aus Stahl (steel screw)
  - Edelstahlschraube | Schraube aus Edelstahl (stainless steel screw)
  - Kupferschraube | Schraube aus Kupfer (copper screw)

- **Application**: preposition: *für* (for)
  - Rigips-Schraube | Schraube für Rigips (screw for plasterboard)

- **Property**: preposition: *mit* (with)
  - Senkkopf-Schraube | Schraube mit Senkkopf (countersunk head screw)

- **Purpose**: preposition: *als/zu* (as/to)
  - Führungsschraube | Schraube als Führung (screw as a guide)
  - Befestigungsschraube | Schraube zu Befestigung (screw as a fixing)
Collecting data for lexicographic work

Principles: implementation is ongoing

- All tool components are applied to the DIY corpus
- Each tool produces
  - result data,
    to be sorted by “central” lexical items: e.g.
      * base of V + N_{OBJ} collocation
      * head of compound
  - For each item of the result data:
    process metadata, to indicate provenience:
      * textual source
      * tool (component) used
- Under way:
  Tool to collect all these data per “central” item and to display it
Collecting data for lexicographic work

Example of a (partial) data collection: s.v. Schraube (screw)

- Adjectives and related compounds and multiword variants:
  - lang – kurz; groß – klein
  - versenkt, seitlich versenkt; Senkkopfschraube
  - metrisch, Schraube mit metrischem Gewinde
  - rostfrei, feuerverzinkt

- Multiword terms with PPs:
  - Schraube mit Sechskantkopf, Sechskantschraube
  - Schraube mit zylindrischem Kopf
  - Gewinde der Schraube

- Verbal contexts:
  - Schraube_{Obj} (ein)drehen, (ein)schrauben; Eindrehen d. S.
  - Schraube_{Obj} anziehen, festziehen
  - Schraube_{Obj} lösen, entfernen
Conclusion

- **Has been shown:**
  - A set of tools to extract terms and their relations
  - Sample results from the DIY domain
  - Proposals for lexicographic use

- **Next steps:**
  - Enhancement of the tools – more detailed evaluations
  - Implementation of further tools needed for work with UGC:
    e.g. coreference resolution, to improve extraction of relations involving verbs and their (pronominalized) complements
  - Implementation of tool to combine the output for lexicographic purposes