Evaluating noise reduction strategies for terminology extraction

Johannes Schäfer, Ina Rösiger, Ulrich Heid, Michael Dorna

Universität Stuttgart, Universität Hildesheim, Robert Bosch GmbH

TIA 2015: Granada, November 2015
Why do we need noise reduction strategies for terminology extraction?
Why do we need noise reduction strategies for terminology extraction?

Standard part-of-speech-based extraction ...

... of N+PP results in:

- Bohrer mit Diamantspitze ✓
  “drill with diamond bit”
- die *Oberfläche mit Leinölfirnis bedecken —
  “cover the *surface with linseed oil varnish”

... of ADJ+N results in:

- sechskantige Schraube ✓
  “hexagonal screw”
- elektromagnetisch *angetriebene Spritzpistole —
  “electromagnetically *operated spray gun”

TIA 2015: Noise reduction strategies for term extraction
Why do we need noise reduction strategies for terminology extraction?

- Standard part-of-speech-based extraction results in
  - Geburtstagskuchen mit Kerzen — “birthday cake with candles”
    → for a technical domain this candidate is not relevant

- Such an expression is typically filtered out by using termhood measures
  - based on a comparison of the domain corpus and a general-language corpus

⇒ But: which termhood measures work well?
⇒ What are their (statistical) properties?
Overview

- Context and objectives
- A standard pipeline approach: components and evaluation
- Improving term extraction quality
  - filtering by syntactic constraints
    - “cover the *surface with linseed oil varnish”
  - filtering out invalid embedded phrases
    - “electromagnetically *operated spray gun”
  - ranking by termhood measures
    - “birthday cake with candles”
- Conclusion and future work
Context and objectives
Context and objectives

Data:

- Expert-produced texts from the DIY domain: manuals, handbooks, articles, ...
- Domain-specific user-generated content (mostly from the web): forums, discussion groups, etc.
  - noisy data
  - requires robust tools
  - part-of-speech based approach better suited than parse-based
Context and objectives

- 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.
  
- Prerequisite: Lexical resources to feed the tools
  - term extraction as a first step
  $\Rightarrow$ quality of extracted terms is important!
A standard pipeline approach: components and evaluation
Data and gold standard

- Domain corpus: collection of texts from the do-it-yourself domain
  - different genres and text types
  - expert and user generated content
  - total number of tokens: 2.7 M
Data and gold standard

- Domain corpus:

<table>
<thead>
<tr>
<th># tokens</th>
<th>text</th>
</tr>
</thead>
<tbody>
<tr>
<td>62,131</td>
<td>do-it-yourself handbook</td>
</tr>
<tr>
<td>6,868</td>
<td>encyclopedia entries</td>
</tr>
<tr>
<td>5,150</td>
<td>list of FAQs with answers</td>
</tr>
<tr>
<td>15,104</td>
<td>tips and tricks for do-it-yourselfers</td>
</tr>
<tr>
<td>35,302</td>
<td>marketing texts</td>
</tr>
<tr>
<td>2,160,008</td>
<td>user generated project descriptions</td>
</tr>
<tr>
<td>444,381</td>
<td>user generated wiki content</td>
</tr>
<tr>
<td>2,728,944</td>
<td>total DIY corpus</td>
</tr>
</tbody>
</table>

- General-language corpus: SdeWaC
  - German web text
  - 880 M tokens

Faß and Eckart 2012
Data and gold standard

- Gold standard
  - 3 independent annotators
  - basic patterns only: N, Adj+N, N+D+N\textsubscript{Gen}, N+P+N
  - decision: [+/- terminologically relevant]
  - we keep track of \{3:0\}-decisions (strict)
    and of \{2:1\}-decisions (liberal)
  - inter-annotator agreement:
    between moderate and substantial agreement
  - contains 4,238 SWTs (including compound nouns) and 826 MWTs
Technology

- Standard corpus technology for preprocessing
  - Tokenizing
  - Tagging, Lemmatization: RF-Tagger
  - Dependency parsing: mate

Schmid 2000
Schmid/Laws 2008
Bohnet 2010, Björkelund et al. 2010
Technology

- Standard term extraction procedures – two steps:
  1) extraction by patterns
  2) ranking of extracted candidates
Technology

- Pattern-based search:
  - POS-shapes:
    - N Kreissäge “circular saw”
    - Adj + N oszillierende Säge “oscillating saw”
    - N + P + N Bohrer mit Kabel “drill with cord”
    - N + D + N_{gen} Spitze des Bohrers “bit of the drill”
  - Allowing for optional determiners, adjectives or adverbs
    - (Adv? Adj? Adj)? N
Improving term extraction quality
Filtering by syntactic constraints
Improving term extraction quality: filtering by syntactic constraints

- Part-of-speech pattern search has no syntactic knowledge
- Problem: candidates covering too long spans
  - typically occur when part of the extracted candidate is actually attached to the verbal phrase

\[
\begin{align*}
die \textit{Schablone mit Farbe besprühen} & \quad \text{“spray the template with paint”} \\
ein \textit{Loch in die Wand bohren} & \quad \text{“drill a hole into the wall”}
\end{align*}
\]
Improving term extraction quality: syntactic validity

Idea:

- Find start and end points of NPs using the dependency parser *mate*  
  Bohnet 2010, Björkelund et al. 2010

- Filtering mechanism:
  - if the POS sequence goes beyond the end point of an NP: invalid
  - else: valid

- Soft filter: only filters out one particular occurrence

- Hard filter: filters out all occurrences of this lemma sequence completely
Improving term extraction quality: syntactic validity

- Terminologically invalid N+PP sequence:

```
VP
   NP  PP
   ein Loch in die Wand
   V   bohren
```

`ein *Loch in die Wand bohren`  "drill a *hole into the wall"
Improving term extraction quality: syntactic validity

- Terminologically valid N+PP sequence:

```
VP
 /   
NP  V
 /  
N   PP
  /  
Bohrer  mit Diamantspitze
```

*Bohrer mit Diamantspitze benutzen*  
*“use drill with diamond bit”*
Improving term extraction quality: syntactic validity

First evaluation:
- 107 N+P+N terms in the gold standard
- Both the hard and the soft filter improve results
- $F_1$-score improvement in figure below
  (performance for ranking by termhood measure $CSmw$)
Improving term extraction quality: syntactic validity

Second evaluation:

- Filter affects more candidates than only N+P+N → also variants of this basic pattern
- 17.4% of all NP+PP candidate occurrences affected
- Precision-based evaluation of top-n lists: 83% for the top 500 candidates

<table>
<thead>
<tr>
<th>Top n</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.75</td>
<td>0.81</td>
<td>0.82</td>
<td>0.82</td>
<td><strong>0.83</strong></td>
</tr>
</tbody>
</table>

Top-n manual plausibility check for “hard” filter
Improving term extraction quality
Filtering out invalid embedded occurrences
Improving term extraction quality: invalid embedded occurrences

- Exhaustive part-of-speech pattern matching
  - not only extracts matches of maximum length
  - extracts matches for all possible patterns

Hartmetallbohrer für faserverstärkte Kunststoffe

“carbide drill for fiber-reinforced plastics”

Hartmetallbohrer

“carbide drill”

faserverstärkte Kunststoffe

“fiber-reinforced plastics”

Kunststoffe

“plastics”
Improving term extraction quality: invalid embedded occurrences

- Exhaustive part-of-speech pattern matching
  - not only extracts matches of maximum length
  - extracts matches for all possible patterns

Hartmetallbohrer für faserverstärkte Kunststoffe

“carbide drill for fiber-reinforced plastics”

Hartmetallbohrer
“carbide drill”

faserverstärkte Kunststoffe
“fiber-reinforced plastics”

Kunststoffe
“plastics”

- Problem: candidates covering too short spans

elektromagnetisch angetriebene Spritzpistole

“electromagnetically operated spray gun”

Spritzpistole
“spray gun”

*angetriebene Spritzpistole
“*operated spray gun”
Improving term extraction quality: invalid embedded occurrences

- C-value handles such nested multi-words
  - Includes term length as number of words
  - Idea: term length including number of elements of compound nouns

<table>
<thead>
<tr>
<th>candidate term</th>
<th>freq</th>
<th>C-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spritzpistole</td>
<td>49</td>
<td>80.00</td>
</tr>
<tr>
<td>*angetriebene Spritzpistole</td>
<td>2</td>
<td>0.00</td>
</tr>
<tr>
<td>elektromagnetisch angetriebene Spritzpistole</td>
<td>2</td>
<td>8.00</td>
</tr>
</tbody>
</table>

Frantzi et al. 2000

- "spray gun"
- "*operated spray gun"
- "electromagnetically operated spray gun"
Improving term extraction quality: invalid embedded occurrences

Extraction of MWTs:

![Graph showing F1-score for top-n candidates (normalized by the length of the GS, N=826).]

- Frequency
- C-value
Improving term extraction quality
Comparing different termhood measures
Improving term extraction quality: termhood measures

- Extracted candidates can be irrelevant for the domain
- Rankings by mere frequency or C-value not satisfactory (F$_1$-score 0.25 and 0.3 respectively)

⇒ Further filtering: top-n of rankings by termhood measures
  - based on a comparison of the term frequency in the domain corpus and a general-language corpus
Improving term extraction quality: termhood measures

- Weirdness ratio for domain specificity (DS)  
  Ahmad et al. 1999

- Corpora-comparing log-likelihood (LL)  
  Rayson and Garside 2000

- Contrastive Selection via Heads (CSvH)  
  Basili et al. 2001

- Term Frequency Inverse Term Frequency (TFITF)  
  Bonin et al. 2010

- Contrastive Selection of multi-word terms (CSmw)  
  Bonin et al. 2010
Improving term extraction quality: termhood measures

Extraction of MWTs:

![Graph showing F1-score for different termhood measures across top-n candidates normalized by the length of the GS, N=826]
Improving term extraction quality: termhood measures

- C-value corrects term frequency with consideration of embeddings in longer terms
- Termhood measures focusing on domain-specificity are mainly based on frequency

⇒ Idea: Use C-value instead of frequency as input for these measures
Improving term extraction quality: termhood measures

Extraction of MWTs:

![Graph showing F1-score for top-n candidates normalized by the length of the GS, N=826]
Improving term extraction quality: termhood measures

Extraction of MWTs:

![Graph showing F1-score vs. top-n candidates normalized by the length of the GS, N=826]

- $F_1$-score
- CSmw(f)
- CSmw(C)

Top-n candidates (normalized by the length of the GS, $N=826$)
Conclusion
Conclusion

Has been shown:

- Part-of-speech based approaches are suitable for term extraction when combined with a set of noise reduction strategies.
- Ensuring syntactic validity helps.
- Filtering out invalid, embedded occurrences improves results.
- Different termhood measures have different statistical properties.
  ⇒ important to make an informed decision.
  ⇒ in our domain and language: DS and CSmw work best.
Future work

Next steps:

- New domain: do measures behave similarly?
- English data: are our results transferable?
- Combination of termhood measures