Universität Stuttgart

BOSCH

Evaluating monolingual term extraction from German texts

Tanja George*, Johannes Schäfer*, Ina Rösiger*, Ulrich Heid•, Michael Dorna◇

*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 (TTC: cf. Gojun et al. 2012)
  - Alternative components: statistical and syntax-based
  - Statistical tool: commercial product (SDL) (SDL MultiTerm 2014 Extract)

Evaluation methodology

- Use of manually designed gold standard:
  - 3 independent annotators: +/- domain specific
  - Patterns:

N	*Bohrmaschine, Schraubenzieher, Loch*
Adj+N	*oszillierende Säge, gebohrtes Loch*
N+$N_{Genitive}$	*Führung der Säge, Kopf einer Schraube*
N+*von*+N	*Fräsen von Kanten, Schleifen von Holz*
N+Prp+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:

Text type	# of tokens	authors
DIY manual	62 131	experts
DIY encyclopedia	6 868	experts
DIY practical "tricks"	15 104	experts
Marketing texts	35 302	experts
DIY project descriptions	2 160 008	UGC
FAQs (forum)	5 150	UGC
Wiki content	444 381	UCG
Total	2 728 944	

Gold standard development

- Guidelines for cases of doubt: term vs. non-term
  - In-domain vs. out-of-domain ambiguities: *Engländer, Rahmen, Leitung, Ton,...*
  - Abbreviations: *PVC*, *EU*
  - Measure indications: *6mm-Bohrer*, *240er Schleifpapier* vs. *2. Gang*, *1-2-do*
  - Product and company names: *IXO von Bosch*
- Size:

Pattern:	{3:0}	{2:1}	**Total**
N	2296	1942	4238
Adj + N	301	303	604
N + N_{gen}	102	46	148
N+*von*+N	42	14	56
N + Prp + N	36	15	51
Total:	2777	2320	**5097**

- Items with f≥4

Inter-annotator-agreement

Annotators:	κ of N+*von*+N:	κ of N+N_{gen}:	κ of N:	κ of ADJ+N:	κ of N+Prp+N:
A1&A2	0.69	0.47	0.50	0.55	0.63
A2&A3	0.65	0.60	0.54	0.54	0.65
A3&A1	0.71	0.48	0.48	0.52	0.60
A1, A2&A3	0.68	0.52	0.51	0.54	0.63

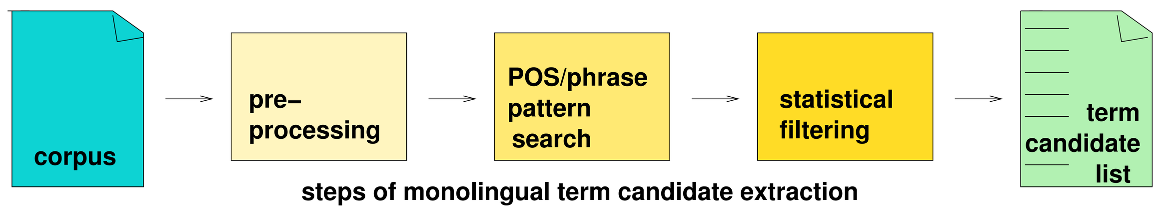
Interpretation of the kappa-values: (Landis et al. 1977)

0.41 - 0.60 Moderate agreement

0.61 - 0.80 Substantial agreement

Tools and components evaluated

1. Hybrid research prototype
   - Text pre-processing: tokenizing, POS-tagging, lemmatization
   - Candidate extraction via POS-patterns
   - Filtering with "weirdness ratio" threshold (Ahmad et al. 1992)



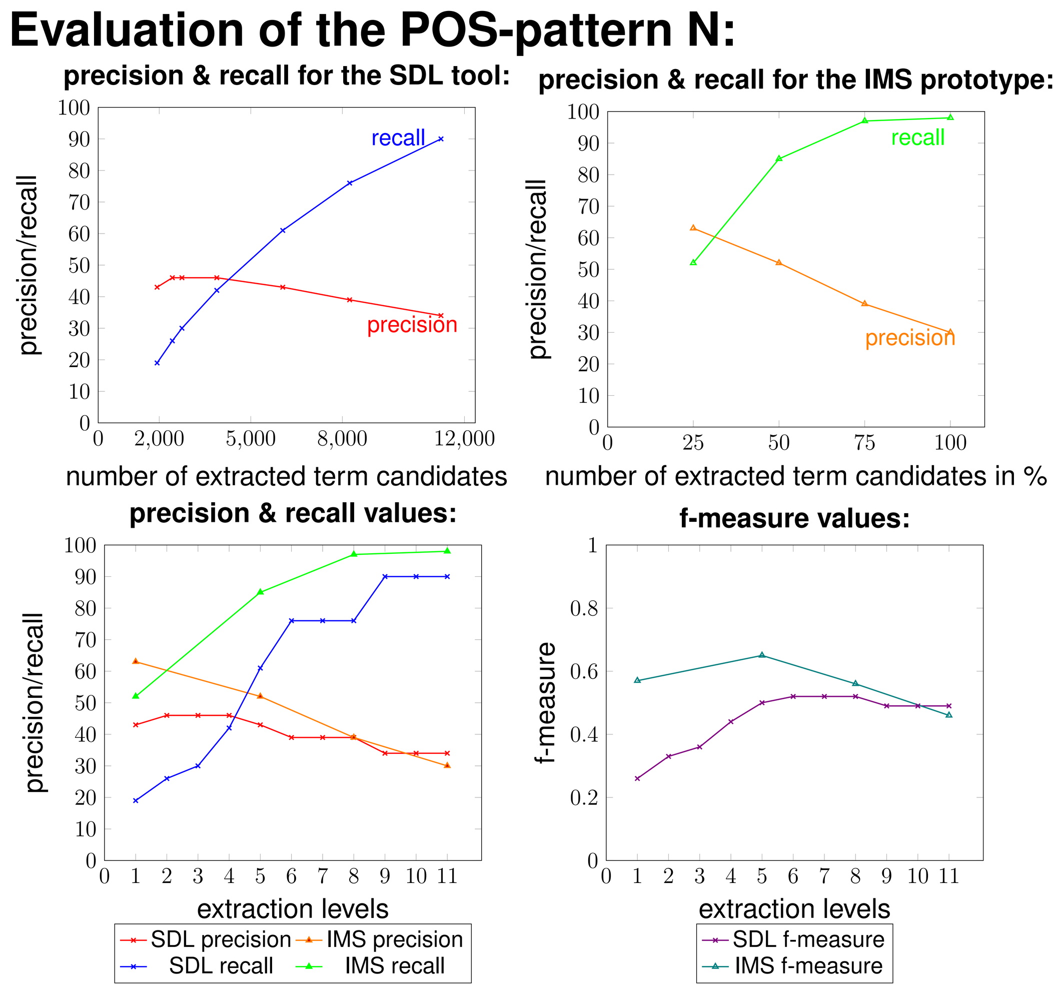
steps of monolingual term candidate extraction

2. Alternative components
   - **Patterns:** NPs and their boundaries, as annotated to subjects/objects by the *mate* parser: check N+Prp+N candidates for phrase boundary compatibility (Bohnet 2010)
   - **Statistical filtering:** Termhood measures (Pazienza et al. 2005)
     - C-value as enhanced frequency (Frantzi et al. 2000)
     - Comparison of domain vs. general language frequency:
       - DS: Domain Specificity (Ahmad 1999)
       - LL: log-likelihood (Rayson/Garside 2000)
       - CSvH: Contrastive Selection via Heads (Basili et al. 2001a)
       - TFITF: Term Frequency Inverse Term Frequency (Bonin et al. 2010)
       - CSmw: Contrastive Selection of multi-word terms (Bonin et al. 2010)
     - First experiments with association measures
- Purely statistical tool (SDL)
  - Language-independent, commercial state-of-the-art
  - 11 user-selectable quality levels with respect to noise ↔ silence relationship

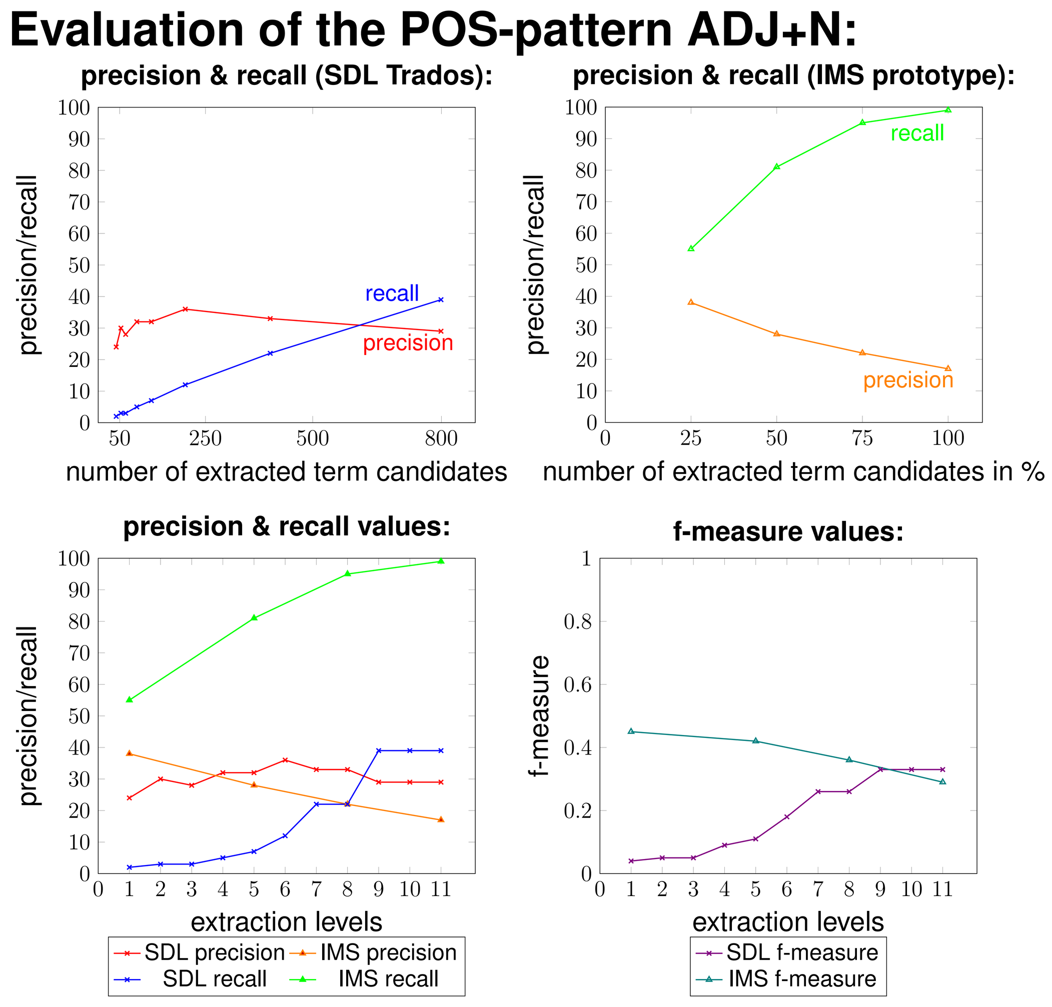
Evaluation 1: Hybrid vs. statistics only

- Tests: (George 2014)
  - Precision and recall vs. length of candidate list (upper panels)
  - Precision and recall vs. SDL quality levels (lower left panel)
  - F-measure by SDL quality levels (lower right panel)

Evaluation of the POS-pattern N:



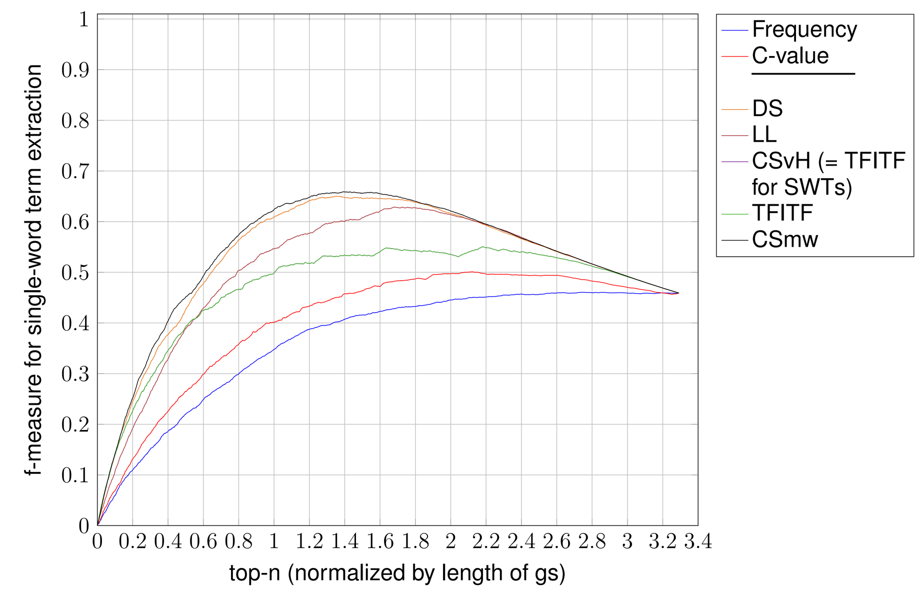
Evaluation of the POS-pattern ADJ+N:



- Results:
  - Hybrid tool outperforms statistical tool for all patterns
  - Several SDL quality levels provide identical results
  - Maximum f-measure: ca. 0.6 for nouns, 0.4 for Adj+N: ⇒ room for improvement

Evaluation 2: Alternative statistical measures

- Tests (Schäfer 2015)
  - Precision, recall, **f-measure** for all patterns and all termhood measures
  - Experiments on combinations of measures



- Results:
  - Weirdness ratio (=DS) and CSmw performed best (max. f-measure: 65% for single-word terms, 50% for multi-word terms (MWTs))
  - C-value "corrects" frequency counts due to sensitivity to term embedding, could be used instead of frequency as input for other measures
  - Association measures outperform frequency baseline, but are only applicable to high frequency candidates

Evaluation 3: Parsing-based extraction

- Motivation: noise in multi-word candidate sets
  - POS patterns: no information about phrase boundaries
  - Example: "NP_1+Prp+NP_2" should only extract NPs, when NP_2 is embedded in NP_1: *Man legt die Oberfräse 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 (George 2014)
  - 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