Acquisition of semantic relations between terms: how far can we get with standard NLP tools?

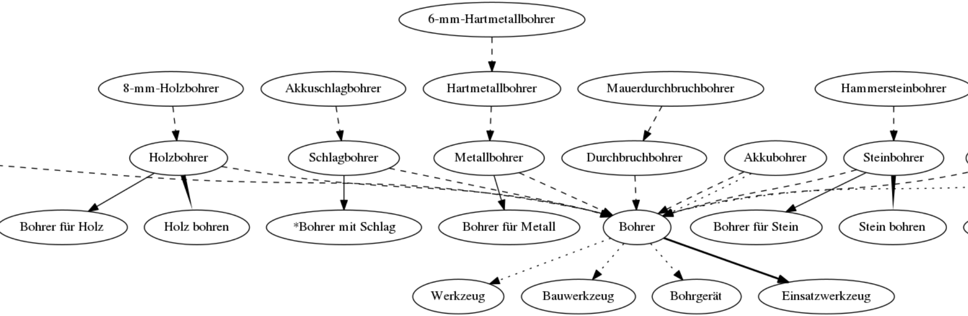
Ina Rösiger, Julia Bettinger, Johannes Schäfer,
Michael Dorna and Ulrich Heid

12 December 2016
5th International Workshop on Computational Terminology
COLING 2016

Aim of this work

- Setting up a detailed data extraction pipeline for the identification and partial classification of terms and their relations
- Checking to which extent relation extraction can be carried out with standard NLP techniques similar to those used in term extraction (without domain adaptation)
- Apply and evaluate these techniques on German user-generated text

Objective



- Extract semantic relations between domain objects
- Evaluate the extraction techniques

Project context

- Project setup:
  - Collaboration, since 10/2014, with
    Robert Bosch GmbH, Corporate Research
- German texts from a broad and heterogeneous domain: descriptions of do-it-yourself (DIY) projects and tools
- Terminology seen in a broad perspective:
  specialized terms plus domain-relevant entities:
  - Not only nominals, but also adjectives and verbs
  - Inclusion of (specialized) collocations
  - Construction of partial hierarchies of domain objects

Overview

Background

- Hybrid term extractor and NLP tools used
- Evaluation methodology

Identifying relational data between terms

- Taxonomic relations
- Non-taxonomic relations

Identifying events involving domain objects

Conclusion and future work

Outline

Background

- Hybrid term extractor and NLP tools used
- Evaluation methodology

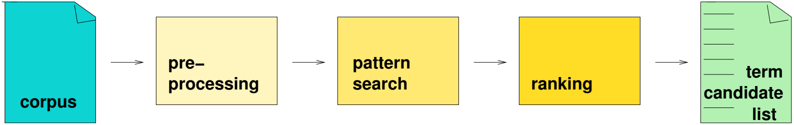
Identifying relational data between terms

- Taxonomic relations
- Non-taxonomic relations

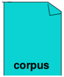
Identifying events involving domain objects

Conclusion and future work

Standard hybrid term extractor



Corpus – text basis



- Heterogeneous data collection:
  - Different DIY-related topics — work with wood and stone, paper, textiles, etc.
  - Different text types:
    - Expert texts (EXP): — DIY encyclopedia, professional project descriptions, etc.
    - User-generated content (UGC): — forum posts by users
  - Different degrees of orality — cf. Koch/Oesterreicher 1985 etc.
- Corpus size: 11M (now: 27M) — EXP ↔ UGC: ca. 1↔5

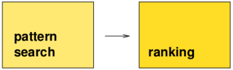
Pre-processing

pre–processing

Use of high-quality tools

- RFTagger: tagging and lemmatisation — Schmid and Laws 2008
- Mate dependency parser — Bohnet 2010
- Morphological analysis: CompoST based on SMOR — Cap 2014, Schmid et al. 2004
- Coreference resolution system — Rösiger and Kuhn, 2016

Pattern search and ranking



Standard hybrid approach

Schäfer et al. 2015

- Part-of-speech patterns to find nominal terms
- (Morpho)-syntactic patterns
  to find predicate-argument structures
- Ranked by termhood measure:
  - comparison with a general-language corpus
  - a set of different measures are implemented

Term candidate list

term candidate list

Nominal term candidates: single word and multi word terms

- Nouns

  Stichsäge, Oberfläche, Bohrung

  jigsaw, surface, drilling

- Adjective+Noun

  doppelseitiges Klebeband, oszillierende Säge

  double-sided adhesive tape, oscillating saw

  vorgebohrtes Loch pre-drilled hole

- More complex patterns

  werkzeugloser Wechsel der Schleifrollen

  tool-free exchange of polishing rolls

Evaluation methodology

- There is no gold standard for relations between domain objects
  → precision-based evaluation only
- Two types of relational data:
  1. Data sorted according to a termhood measure:
     - "Good terms" at the top of the list
     - "Non-terms" at the end of the list
     - → Mainly top of lists to be evaluated,
       as non-terms will be excluded from further analysis
  2. Data sorted according to token frequency in corpus
     - Frequent items at the top of the list
     - Rare items at the end of the list
     - → Frequent items more relevant for quality assessment
       of extraction results: ⇒ stop evaluation at
       e.g. $f = 10$

Outline

Background

Hybrid term extractor and NLP tools used

Evaluation methodology

Identifying relational data between terms

Taxonomic relations

Non-taxonomic relations

Identifying events involving domain objects

Conclusion and future work

Identifying relational data between terms

Relational data: Taxonomic (= subtype) relations

- Two techniques
  - Definition-like patterns (cf. Hearst 1992)

- "Eine Vertikalbandsäge ist eine Säge, die ..."
"A vertical band saw is a saw which ..."
- "Vertikalbandsägen gehören zur Gruppe der Bandsägen."
"Vertical band saws belong to the group of band saws."

  - Morphological analysis (see paper for details)

- Säge	*saw*
– Bandsäge	*band saw*
— Elektrobandsäge	*electrical band saw*
— Hand-Bandsäge	*manual band saw*
— Horizontalbandsäge	*horizontal band saw*
— Vertikalbandsäge	*vertical band saw*

→ evaluation of tools ongoing

Taxonomic relations: Hearst patterns

Extracting hyponymy pairs from ...

- Definition-like sentences (“an X is a Y which ...”) and from list-like enumerations (“Xs, such as Y1, Y2 ...”) Hearst 1992
  - Nominal patterns
    on the basis of pos and lemma sequences
  - Verbal patterns:
    extracted from parsed text by use of verbal predicates which denote class membership

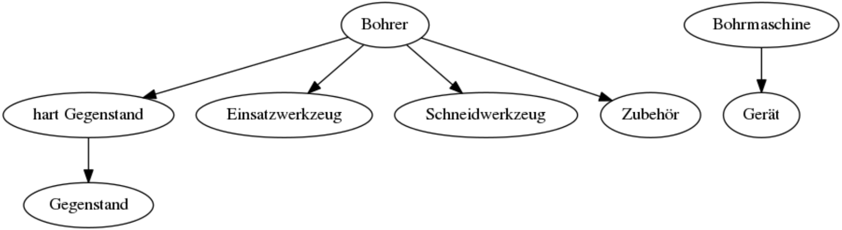
gehören zu *belong to*,
zählen zu *be part of*

Taxonomic relations: Hearst patterns

Implementation:

- German version of the classical hypernym patterns: not mere translations from English, carefully adapted with many constraints on the pos and lemma level
- Four main patterns:
  - N_{sub1} , N_{sub2} (und|oder) (ander.*|vergleichbar.*|sonstig.*|weiter.*) (Adj)? N_{sup}
  - (Adj)? N_{sup} (,)? insbesondere (Adj)? N_{sub}
  - (Adj)? N_{sup} (,)? einschließlich (Adj)? N_{sub}
  - (Adi) N_{sup}wie N_{sub1} (,)? N_{sub2} ((’und|oder|sowie’) (Adj) N_{sub3}))*

Taxonomic relations: Hearst patterns



- A subset of relations found for exemplary term *Bohrer (drill)* using Hearst patterns
- Arrows indicate a relation of hyponymy,
  e.g. "*Bohrer* is-a *Schneidewerkzeug*"

Taxonomic relations: Hearst patterns

Evaluation:

- Type 2 evaluation: based on frequency
- First evaluation:
  - Top 200 search result pairs sorted by frequency
  - Decision: does the hyponymy relation hold?
  - True for 163 out of the 200 pairs 82%
- Second evaluation:
  - Pairs are filtered out in which none of the two nouns is a term
  - Remaining pairs sorted by frequency
  - Two-fold evaluation:
    - validity of the hyponymy relation: 164/200 82%
    - domain relevance: 151 out of the 164 valid pairs 92%

Non-taxonomic relations: compounds and their paraphrases

- Many compound terms are paraphrased as NP+PP constructions
- Preposition makes the relation explicit
  which exists between the compound and its modifier
  e.g. material: *Stahlschraube* ↔ *Schraube aus Stahl* steel screw
- The same holds for complex NPs
  Holz der Fichte ↔ *Holz aus Fichte* ↔ *Fichtenholz* spruce wood
- The most frequent paraphrase tends to be the adequate one
- Prepositions may be ambiguous:
  issue less acute within our discourse domain

Non-taxonomic relations

- Material: *Stahlschraube ↔ Schraube aus Stahl*

  steel screw – screw made of steel

- Property: *Senkkopfschraube ↔ Schraube mit Senkkopf*

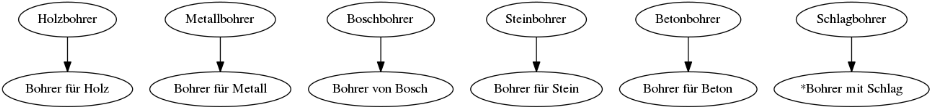
  countersunk screw – screw with countersunk head

- Purpose: *Führungsschraube ↔ Schraube als Führung*

  guide screw – screw as guide

Compound	Paraphrase	Relation
Steinbohrer (stone drill)	Bohrer für Stein (for)	purpose
Metallbohrer (metal drill)	Bohrer für Metall (for)	purpose
Schutzfolie (protection film)	Folie zum Schutz (for)	purpose
Diamantbohrer (diamond drill)	Bohrer aus Diamant (made of)	material
Aluprofil (aluminium profile)	Profil aus Alu (made of)	material
Heizkörperverkleidung (radiator cover)	Verkleidung vor Heizung (in front of)	location
Kellerraum (basement room)	Raum im Keller (in)	location

Non-taxonomic relations: compounds and their paraphrases



- A subset of relations found for exemplary term *Bohrer (drill)* by matching compounds and their NP+PP paraphrases
- Arrows indicate different semantic relations, depending on the preposition

Non-taxonomic relations: compounds and their paraphrases

Evaluation:

- Type 2: based on frequency
- First evaluation:
  - Top 200 paraphrase-compound pairs, sorted by compound frequency
  - Decision: valid paraphrase?
  - 157 out of 200 paraphrases are valid 79% type accuracy
  - Errors are mainly due to implausible prepositions, such as *Rest im Holz (rest in the wood)* for *Holzrest* (*scrap wood)*

Non-taxonomic relations: compounds and their paraphrases

- Second evaluation: until f=20, plus 12 ≥ f ≥ 10
- 1224 out of 1737 pairs are good: 70.4%
- Prepositions which never produced relevant paraphrases: *außer, bezüglich, bis, hinter, je, ohne, per, trotz, etc.*:
- Certain prepositions provide good results:
  - *aus*: Material: 92.7%, *für*: Purpose: 87.4%
- Other prepositions provide a mixed result: *als*: 65%, *an*: 52%
- Expectably, genitive paraphrases plus *von-PPs* of *-ung*-compounds are good: 105 of 111 cases 94.6%
- Concentrating on genitives and best suited prepositions, i.e. *aus, für, gegen, von, vor* 88.6% (N = 1069)

Outline

Background

- Hybrid term extractor and NLP tools used
- Evaluation methodology

Identifying relational data between terms

- Taxonomic relations
- Non-taxonomic relations

Identifying events involving domain objects

Conclusion and future work

Predicate argument structures

- Verb object pairs:

  Holz bohren (to drill wood), *einen Kreis bohren (to drill a circle)*, ...
- Subject verb pairs:

  Holz verzieht sich (wood warps),

  eine Absaugeeinrichtung spart Zeit (a suction device saves time)
- Verb-dependent and adjunct PPs:

  auf Gehrung sägen (to miter), *für Stabilität sorgen (to ensure stability)*,

  mit der Stichsäge ausschneiden (to cut with a jigsaw)
- Predicative constructions:

  Bohrer ist ein Elektrowerkzeug (drill is a power tool)

  Spitze ist besonders dünn (tip is very thin)
- Negation:

  die Sicherheitskappe nicht abziehen (do not remove the safety cap)
- Adverbs:

  heiß verleimen (to hot glue), *trocken reiben (to rub dry)*,

  dünn beschichten (to coat thinly)

Predicate argument structures

Implementation:

- Extraction is based on dependency parser `mate`
- We can either extract whole phrases
  or just the heads of the phrases
- Extractors can be combined to search for longer patterns, e.g.
  - *Holzspiralbohrer haben eine lange Zentrierspitze*

    wood drills have long lathe centers
  - *Beton besteht aus Zement und Wasser...*

    concrete is made of cement and water
  - *Kupfer benötigt keinen schützenden Anstrich*

    copper requires no protective coat

Predicate argument structures

Evaluation: Verb object pairs

- Type 1: sorted by termhood measure “weirdness ratio”

  Ahmad et al. 1992

- Top 250 pairs
- First evaluation:
  - Decision: syntactically valid, given example sentence
  - 15 out of 250 are invalid 94% accuracy
  - Parsing quality

  seems well-suited as a basis to extract data

Predicate argument structures

Evaluation: Verb object pairs

- Second evaluation (1/2):
  - Decision:
    domain relevance: term, no-term, preprocessing error
  - Excluding three verbs: *haben, sein, geben* *(have, be, give)*
  - 27 out of 250 (10%) were preprocessing errors
  - 150 out of 250 (60%) good terms
  - 73 out of 250 (30%) bad terms

Predicate argument structures

Evaluation: Verb object pairs

- Second evaluation (2/2): error analysis
  - Bad terms: often, extraction pattern does not cover verb subcategorization in full:

    Werfen Sie Elektrowerkzeuge nicht in den Hausmüll

    ⇒*Elektrowerkzeuge*$_{OBJ}$+*werfen* $_V$

    (Do not throw power tools into the trash ⇒ *throw*$_V$+ *power tools*$_{OBJ}$*)*

Predicate argument structures

Evaluation: Verb p-object pairs

- Type 1 evaluation: sorted by termhood measure "weirdness ratio" Ahmad et al. 1992
- Top 200 precision-based evaluation
- Decision: syntactically valid, given an example sentence?
- 191 out of 200 are syntactically plausible 96%
- Most of the extracted pairs are domain relevant

  für festen Halt sorgen (ensure stability), zum Lieferumfang gehören (belong to delivered items), auf Gehrung sägen (to miter), mit Kies beschweren (weigh down with gravel), auf Rechtwinkligkeit achten (ensure perpendicularity).
- Almost all bad pairs: PP attachment problems

  [suchen mit Akkubetrieb]: Ich suche ein Gerät mit Akkubetrieb

  (I'm looking for a device with battery operation)

The role of coreference resolution ...

... for the enhancement of recall
in the extraction of predicate argument structures

- Idea: replace
  pronominalised arguments with proper or common NPs
- Affects 40% of objects in extracted verb object pairs
- Indirect evaluation of coreference tool: Rösiger and Kuhn, 2016
  how do the extracted verb object pairs change?

The role of coreference resolution ...

... for the enhancement of recall
in the extraction of predicate argument structures

- Indirect evaluation:
  - More candidate pairs: 5% more candidates
  - Newly found candidate pairs: good candidates?
    82% of the 193 new candidates relevant to domain
    120er-Schleifpapier verwenden (use 120-grit sandpaper),
    6-mm-Loch bohren (drill 6-mm hole).
  - More evidence for pairs already retrieved:
    higher frequencies

Verb-derived items as a source of relational data

- Application of verb predicate pairs
- For compounds with nominalized verbs as heads: search for verbs and their object as the non-head of the compound
- If we find a paraphrase: evidence that the compound describes an event corresponding to the verb and its object

Compound	Paraphrase
Bodendämmung (floor insulation)	Boden dämmen (insulate floors)
Fensterisolierung (window insulation)	Fenster isolieren (insulate windows)
Betonbohrung (concrete drilling)	Beton bohren (drill concrete)
Leimverteilung (paste distribution)	Leim verteilen (distribute paste)

Verb-derived items as a source of relational data

Evaluation:

- Type 2: sorted by frequency
- Top 125 + bottom 125 compounds sorted by frequency
- Decision: Valid paraphrase for a given compound?
- Top 125: 74% valid
- Bottom 125: 82% valid

Outline

Background

- Hybrid term extractor and NLP tools used
- Evaluation methodology

Identifying relational data between terms

- Taxonomic relations
- Non-taxonomic relations

Identifying events involving domain objects

Conclusion and future work

Conclusion

We have shown...

- A set of techniques used
  to acquire semantic relations between terms
  (without domain adaptation)
- Overall: acceptable precision
  when applying standard NLP tools to relation extraction

Future work

- Integrate more morpho-syntactic patterns
  to extract more relations
- Include other approaches to extract relations, e.g.
  distributional methods to distinguish semantic relations
- Evaluation: more detailed
  - also for infrequent cases
  - more annotators, to be able to assess human agreement

Thank you!

Questions?

✉ ina.roesiger@ims.uni-stuttgart.de
heid@uni-hildesheim.de