Term variation as a source of relational knowledge

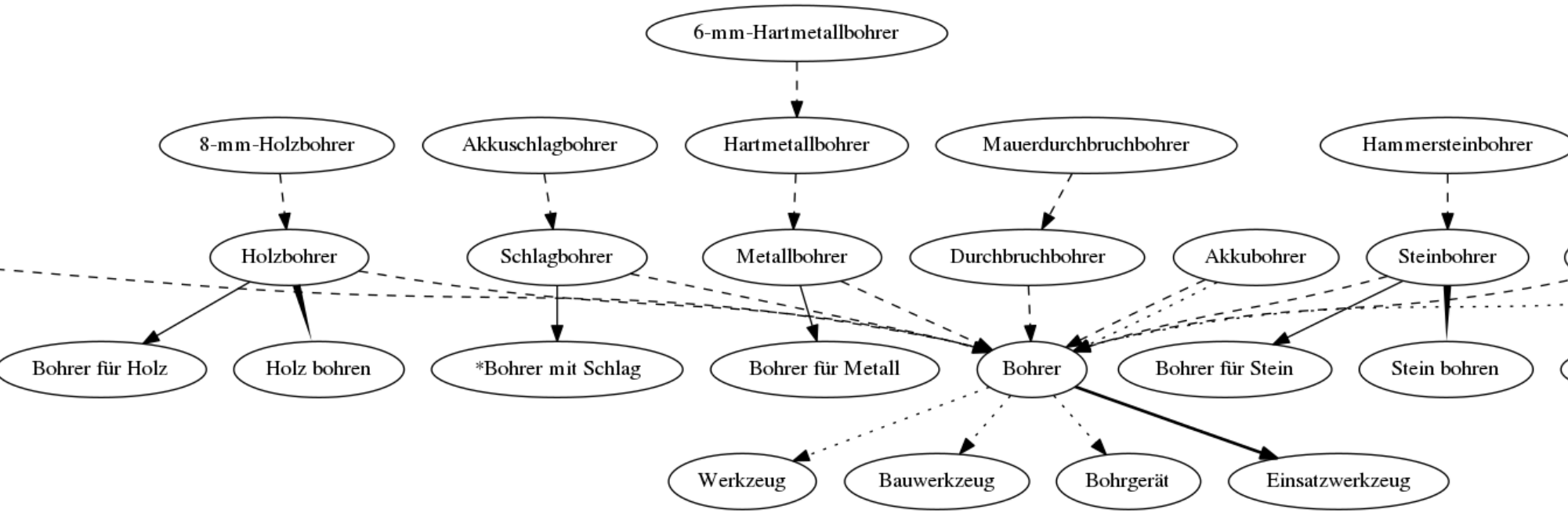
Ina Rösiger, Johannes Schäfer,
Michael Dorna, Anna Hätty and Ulrich Heid

21 July 2017
TermVar: Workshop on Terminological Variation,
Hildesheim

Aim of this work

- A detailed data extraction pipeline for the identification
  (1) of terms (= term extraction)
  (2) of semantic relations between the terms
  (3) of events involving the terms
- Focus in this talk on (2) and (3):
  Term variation as a source of relational knowledge
- Apply and evaluate these techniques
  on German user-generated text

Objective: relational knowledge



- Based on term variation:
  extract semantic relations between domain objects
- Evaluate the extraction techniques

Objective: relational knowledge

Based on two variant types:

(1) Compounds (N+N) and NP PP variants:

Metallbohrer – Bohrer für Metall *metal drill - drill for metal*

→ we can derive the semantic relation between the two compound parts: *purpose*

(2) Compounds ($N+N_{deverbal}$) and verb-object variants:

Holzbohrer – Holz bohren *wood drill - to drill wood*

→ we can assume an event reading

Project context

- Project setup:
  - Collaboration, since 10/2014, with
    Robert Bosch GmbH, Corporate Research
    (Dr. Michael Dorna and Anna Hätty)
- 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 (specialised) collocations
  - Construction of partial hierarchies of domain objects

Overview

Background

- Hybrid term extractor and NLP tools used
- Evaluation methodology

Term variation as a source for relational knowledge

- Semantic relations
- Events

Conclusion

Outline

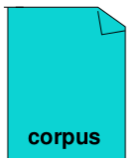
Background
- Hybrid term extractor and NLP tools used
- Evaluation methodology

Term variation as a source for relational knowledge
- Semantic relations
- Events

Conclusion

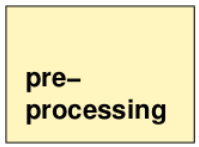
Standard hybrid term extractor

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
- Corpus size: 11M — EXP ↔ UGC: ca. 1↔5

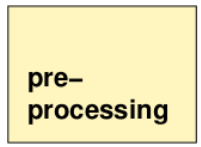
Pre-processing



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

Pre-processing



Use of high-quality tools

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

CompoST: morphological analysis of German

- We split compounds
  using the splitting tool CompoST Cap 2014
- Implementation is aware of complex non-heads, i.e.
  we check for attested morpheme combinations in our corpora
  to exclude wrong splits.
- For example, for *Eigenbaubandsäge ("self-constructed bandsaw")*,
  - first split into morphemes (Eigen | bau | band | säge)
  - then check for attested combinations:
    Bandsäge (valid, found), *Baubandsäge* (not found),
    Eigenbau-X (valid, found),
    resulting in the correct split *Eigenbau* | *Bandsäge*.

CompoST: morphological analysis of German

- Compound analysis: determinative compounds
  can be interpreted as hyponyms of their morphological heads

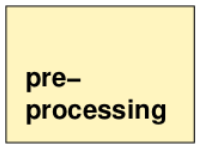
  Band|säge → Säge band|saw→ saw.

- Head as the superordinate
  and compounds as subtypes of their heads:
  Säge (saw) has subordinates such as
  Kreissäge (circular saw), *Bandsäge (bandsaw)*.
- We build a partial hierarchical structure for every head

- *Säge*
  - *Bandsäge*
    - *Elektrobandsäge*
    - *Hand-Bandsäge*
    - *Horizontalbandsäge*
    - *Vertikalbandsäge*

saw
- *band saw*
  - *electrical band saw*
  - *manual band saw*
  - *horizontal band saw*
  - *vertical band saw*

Pre-processing

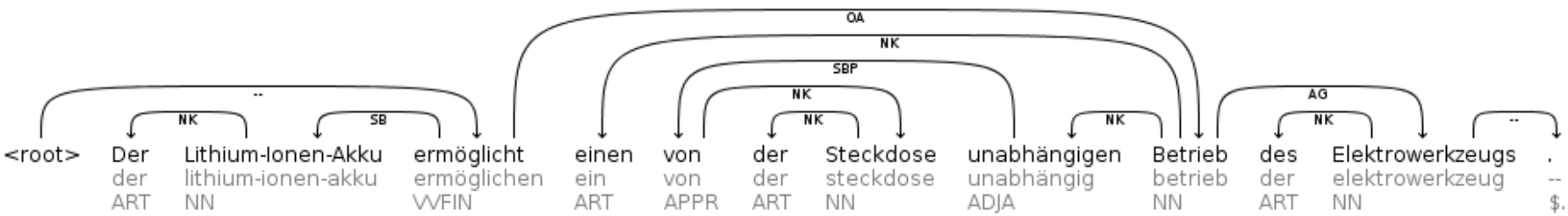


Use of high-quality tools

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

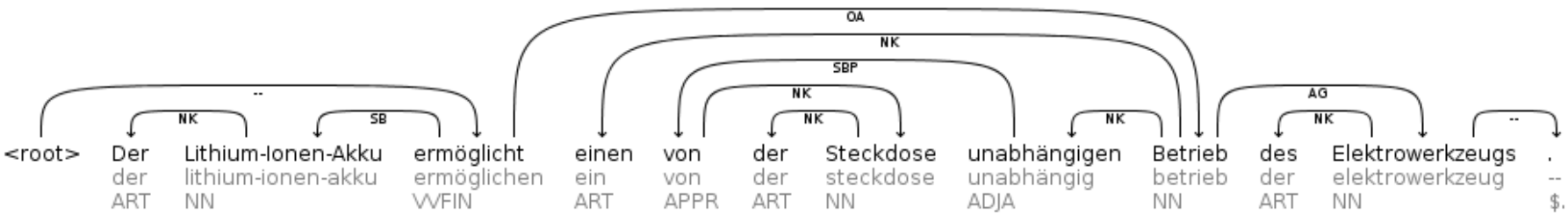
Dependency parsing

- Extraction is based on dependency parser `mate`



- 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"

Dependency parsing and script



0	Der	SUBJ-Embedded	The
1	Lithium-Ionen-Akku	SUBJ-Head	lithium ion accumulator
2	ermöglicht	VERB-Active	enables
3	einen	OBJ-Embedded	a
4	von	OBJ-Embedded	from
5	der	OBJ-Embedded	the
6	Steckdose	OBJ-Embedded	socket
7	unabhängigen	OBJ-Embedded	independent
8	Betrieb	OBJ-Head	operation
9	des	OBJ-Embedded	of the
10	Elektrowerkzeugs	OBJ-Embedded	power tool
11	.	NULL	.

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:

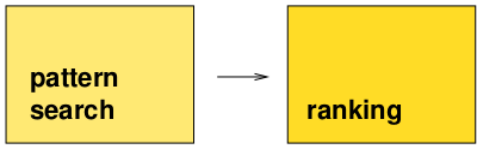
  heiß verleimen (to hot glue), *trocken reiben (to rub dry)*,

  dünn beschichten (to coat thinly)

Extraction of predicate argument structures

- 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

Pattern search and ranking

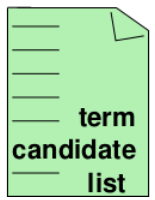


Standard hybrid approach

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



- Nominal term candidates:

  Stichsäge, doppelseitiges Klebeband, werkzeugloser Wechsel der Schleifrollen

  jigsaw, double-sided adhesive tape, tool-free exchange of polishing rolls

- Predicate-argument structures:

  Loch bohren, für Stabilität sorgen

  to drill a hole, to ensure stability

→ Basis for variant and relation extraction

Relation extraction

- Taxonomic/subtype relations:
  - Morphological analysis using CompoST:

    Band|säge → Säge band|saw→ saw.
  - Definition-like (Hearst) patterns:

    "Eine Vertikalbandsäge ist eine Säge, die ..."

    "A vertical band saw is a saw which ..."
- Non-taxonomic relations:
  - Compounds and their NP PP variants
  - Deverbal compounds and their verb+object variants
  - (Predicate argument structures)

Evaluation methodology

- Gold standard only for nominal candidates
- For relations and events: → precision-based evaluation only
- Two types of relational data:
  (1) Data sorted according to a termhood measure:
  (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. $f = 10$
- All evaluations in this talk are of type (2)

Outline

Background
Hybrid term extractor and NLP tools used
Evaluation methodology

Term variation as a source for relational knowledge
Semantic relations
Events

Conclusion

Term variation as a source for relational knowledge: Semantic relations

- 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
  (when several PP alternatives exist)
- Prepositions may be ambiguous:
  issue less acute within our discourse domain

Term variation as a source for relational knowledge: Semantic relations

- Material: *Stahlschraube* ↔ *Schraube aus Stahl*

  steel screw – screw made of steel

- Property: *Senkkopfschraube* ↔ *Schraube mit Senkkopf*

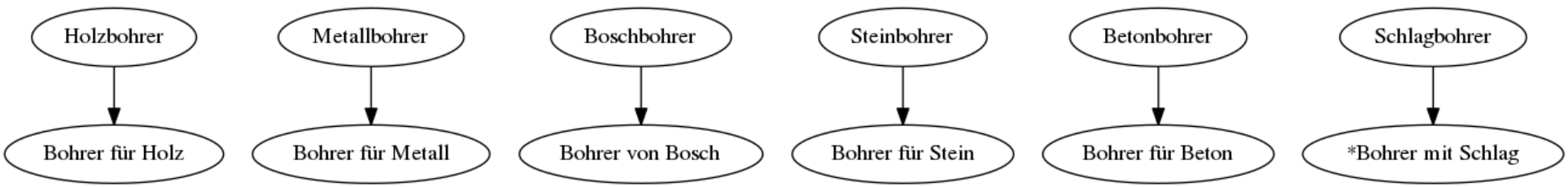
  countersunk screw – screw with countersunk head

- 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

Term variation as a source for relational knowledge: Semantic relations



- 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

Term variation as a source for relational knowledge: Semantic relations

Evaluation:

- First evaluation (based on frequency):
  - 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 (leftover in the wood)* for *Holzrest* (*scrap wood)*

Term variation as a source for relational knowledge: Semantic relations

- Second evaluation: until f=20, plus $12 \geq f \geq 10$
- 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%

Term variation as a source for relational knowledge: Semantic relations

- 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)

Term variation as a source for relational knowledge: Semantic relations

Tool refinement

- Collapsing genitives and *von*-PPs into one category of paraphrase
- Possibly excluding certain prepositions from paraphrase tests
- Possibly running paraphrase extraction not on POS-shapes, but on parsed data, as some paraphrase candidates are not syntactically well-formed

Term variation as a source for relational knowledge: Events

- Compounds and verb-object variants:

  Holzbohrer – Holz bohren *wood drill - to drill wood*

  → we can assume an event reading
- For compounds with nominalised 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)

Term variation as a source for relational knowledge: Events

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

Term variation as a source for relational knowledge
- Semantic relations
- Events

Conclusion

Conclusion

We have shown...

- that term variants are a source for relational knowledge
  - NP PP variants of compound phrases make the relation between the compound and its modifier explicit
  - Verb object variants of deverbal events suggest event reading
- that they can be extracted with acceptable precision when applying standard NLP tools

Thank you!

Questions?

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