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

Ina Rösiger[1], Julia Bettinger[1], Johannes Schäfer[1], Michael Dorna[2] and Ulrich Heid[3]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Robert Bosch GmbH, Germany, [3]University of Hildesheim, Germany
{roesigia|bettinja|schaefjs}@ims.uni-stuttgart.de
michael.dorna@de.bosch.com, heid@uni-hildesheim.de

Abstract

The extraction of data exemplifying relations between terms can make use, at least to a large extent, of techniques that are similar to those used in standard hybrid term candidate extraction, namely basic corpus analysis tools (e.g. tagging, lemmatization, parsing), as well as morphological analysis of complex words (compounds and derived items). In this article, we discuss the use of such techniques for the extraction of raw material for a description of relations between terms, and we provide internal evaluation data for the devices developed.

We claim that user-generated content is a rich source of term variation through paraphrasing and reformulation, and that these provide relational data at the same time as term variants. Germanic languages with their rich word formation morphology may be particularly good candidates for the approach advocated here.

1 Introduction

While term candidate extraction from texts typically targets domain objects, a fuller domain model, as needed for terminological, lexicographic or text classification purposes, requires in addition the provision of data on hyponymy relations between domain objects (taxonomic relations), on properties of domain objects and on events that involve these domain objects.

The objective of this paper is to provide an assessment of the applicability of standard state-of-the-art computational linguistic tools for the task of extracting evidence from which taxonomic relations between domain objects, as well as events involving the domain objects can be derived. We work with German data, but we expect most of our results to be generalizable to other Germanic languages. The tools in question are (i) basic corpus preprocessing tools (tokenizing, pos-tagging, lemmatization, parsing) as well as coreference resolution, (ii) query tools applicable to the preprocessed corpora and (iii) word formation analyzers. We use these tools, because we also carry out term candidate extraction on the basis of this same infrastructure and intend to explore to which degree one and the same standard hybrid approach can be used both to extract term candidates and to extract evidence for relations between them. In this paper, we do not address actual ontology construction.

The remainder of this paper is structured as follows: Section 2 presents the background of our experiments: the text collection used, as well as the tools for pre-processing, data extraction and ranking of term candidates. In Section 3, we discuss the extraction of evidence for relations between domain objects, in terms of relevant linguistic phenomena, different extraction techniques and, for each one, first evaluation results. Section 4 is structured in parallel to Section 3 and deals with raw material for verb-derived events involving domain objects. A comparison with the state of the art follows in Section 5 and we conclude in Section 6.

2 Background and objectives

2.1 Text basis

We use a corpus of German forum posts collected from several online forums in the domain of do-it-yourself (DIY) projects, e.g. work with wood or stone. The posts have been contributed in part by domain experts (giving e.g. advice on techniques, tools, etc.) and in part by end users describing their own projects[1]. Alongside, we use texts from a few professional sources, such as an online encyclopedia and a wiki for DIY work, tools and techniques. The corpus used for the work described here totals ca. 11 M words, with 20% expert text vs. 80% end-user data.

Forum data, as most user-generated content, presents properties of orality (in the sense of Koch and Oesterreicher (1985)): greeting forms (*hallo, tschüss*), contracted forms (verb+pronoun: *hamse* for *haben sie* etc.), orthographic, morphological and syntactic deviance. We also find elements typical of computer-mediated communication, such as addressing (@Peter: ...) or emoticons. The texts contain deviant orthography, spelling errors, compounds written in two chunks instead of one (*Bohrer Spitze* for *Bohrerspitze*, drill bit) etc., covered partly by normalization at tokenizing time. We cannot yet quantify the loss in recall due to these deviances, as far as e.g. parsing-based data extraction is concerned (cf. however Section 4.2.1 and 4.2.2 for precision figures). Terminology in these texts is characterized (i) by term variation ((morpho-) syntactic, in Daille (2007)'s terms) and (ii) by considerable amounts of specialized terms also retrievable from conceptually oral texts[2].

2.2 A standard hybrid term candidate extractor and its computational linguistic components

The extraction of relations between terms presupposes a preceding step of term candidate extraction. Our system uses a standard hybrid approach (cf. Schäfer et al. (2015)): on the basis of either tagged and lemmatized or of parsed text ("preprocessing" in Figure 1), it first applies symbolic patterns (pos-patterns or (morpho-)syntactic patterns) to extract all candidates that follow a given pattern ("pattern search" in Figure 1), before computing termhood measures (such as Ahmad et al. (1992)'s weirdness ratio) to rank candidates by comparison with a general-language corpus (SdeWaC (Faaß and Eckart, 2013)). In the standard term candidate extraction mode, domain experts are then asked to verify the term candidates. Variant recognition is an optional part of the same architecture.

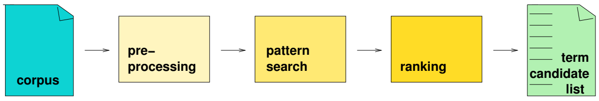

Figure 1: Steps in term candidate extraction: overview

The texts are tokenized and normalized (homogeneous orthography of e.g. numeric indications, cf. *60x40 cm*), tagged and lemmatized using RFTagger (Schmid and Laws, 2008), and dependency parsed using the mate parser (Bohnet, 2010). An automatic correction step is applied for lemmatization. Dependency parses are in addition annotated with phrase boundaries and heads, such that information corresponding to both techniques, constituent and dependency parsing, is available: the full verb of each sentence, its subject and complements, as well as adjuncts and negation are annotated and thus retrievable as context parameters.

An additional step of linguistic annotation is coreference resolution and discourse processing. We use IMS HotCoref DE (Rösiger and Kuhn, 2016), a state-of-the-art coreference resolver for German. In a post-processing step, we annotate personal, possessive, demonstrative and relative pronouns with the closest non-pronominal antecedent identified by the resolver. Experiments on the use of coreference resolution to enhance recall in the extraction of verbs and their arguments can be found in Section 4.2.3.

For compound splitting we use CompoST (Cap, 2014), a compound splitter which combines the use of a rule-based morphology system (SMOR (Schmid et al., 2004)) with morpheme verification in corpus data, thereby extending and improving on the approach proposed by Koehn and Knight (2003) for statistical machine translation. For all components of a compound, including those which are complex themselves, the tool verifies the presence and number of occurrences in a (set of) texts. In our application, the do-it-yourself corpus is used as a knowledge source for this check, in addition to a (newspaper-based) general language corpus. Splits that involve implausible or rare components are dispreferred.

[1] A typical forum of this type is "1-2-do.com"

[2] Work on quantifying the terminological richness of more vs. less oral/CMC texts is under way.

Pattern-based search on all levels, with the exception of coreference resolution, is performed by use of the Corpus Workbench (CWB) system (Evert and Hardie, 2011).

2.3 Objective: Assessment of applicability for the extraction of evidence for relations

The architecture and tools described above may be combined to support the search and retrieval of evidence for relations between objects and for events. The objective for the present article is to provide an assessment of the precision of the standard tools when applied to relation extraction. An assessment of recall requires the availability of gold standard data; while work on manual annotation of relations is ongoing, this resource is not yet complete.

Figure 2 shows the collection of semantic relations for the exemplary term *Bohrer (drill)*. The different arrows represent the source of the semantic relation as well as its type. The remainder of the paper will present the techniques used and evaluations of these techniques.

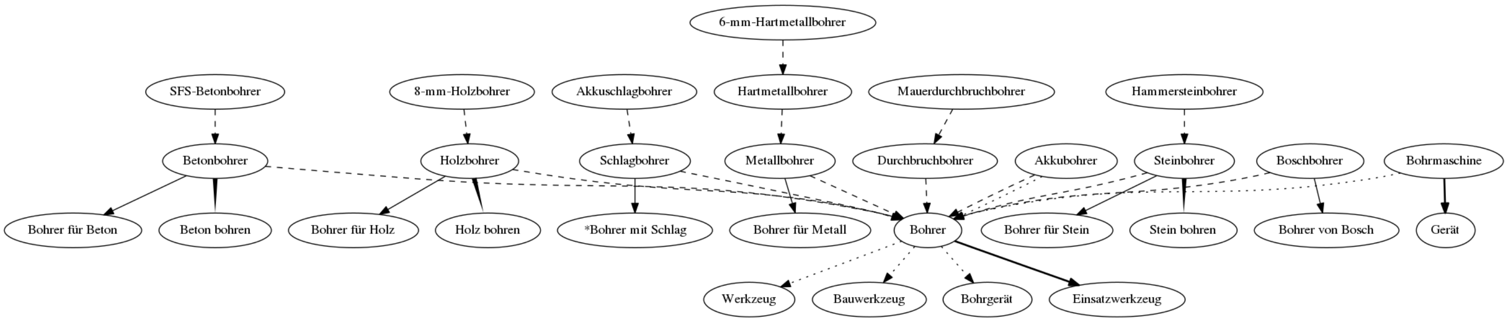

Figure 2: An exemplary subset of relations found for the term *Bohrer*. Bold lines = Hearst patterns (hyponymy relation), normal lines= compounds and their nominal paraphrases (synonymy), dashed lines= compound analysis (hyponymy), broad lines= compounds and their verbal paraphrases (associated events), dotted lines= GermaNet (hyponymy). Not included due to space restrictions are verbs and their arguments.

3 Identifying relations between domain objects

3.1 Relevant phenomena

Taxonomic relations between domain objects: Taxonomic (= hyponymy) relations can be extracted from definition-like sentences (“an X is a Y which ...”) and from list-like enumerations (“Xs, such as Y1, Y2 ...”), as first discussed for English by Hearst (1992). Such relations may also be extracted from parsed text by use of verbal predicates which denote class membership (e.g. *gehören zu* (“belong to”), *zählen zu* (“be part of”) etc.).

Similarly, determinative compounds can be interpreted as hyponyms of their morphological heads (*Band|säge → Säge*, *“band|saw”→ “saw”*).

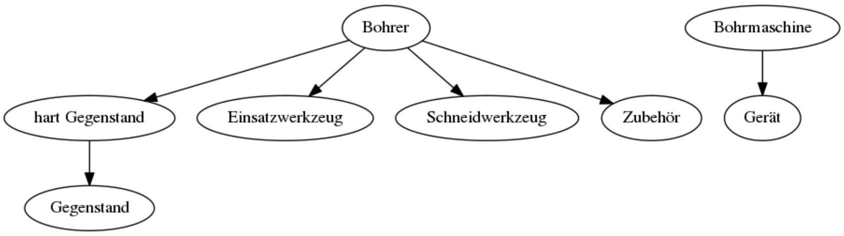

Figure 3: A subset of relations found for *Bohrer* using Hearst patterns; arrows indicate a relation of hyponymy, e.g. “*Bohrer* is-a *Schneidewerkzeug*”.

Figures 3 and 4 show an exemplary subset of taxonomic relations for the term *Bohrer* (drill). The figures show partial hierarchies derived from result data of each procedure. As Figure 4 shows, no

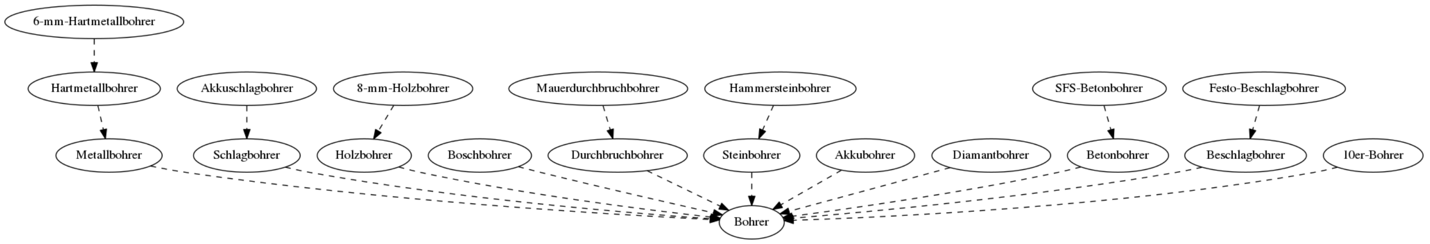

Figure 4: A subset of relations found for *Bohrer* by compound analysis; arrows indicate a relation of hyponymy, e.g. "*Holzbohrer* is-a *Bohrer*".

inferencing or synonym search has yet been applied (we consider such techniques to be part of the actual ontology construction work), so that e.g. *10er-Bohrer* and *10-mm-Bohrer* are not identified as synonymous, and *Akkubohrer* is not related with *Akkuschlagbohrer*.

Non-taxonomic relations between domain objects: In our texts, many compound terms are paraphrased by means of NP+PP constructions where the preposition makes the relation explicit which exists between the compound head and its modifier. Obviously, prepositions themselves may be ambiguous, in unrestricted contexts, with respect to the relation they indicate; this problem is however less acute within the discourse domain of DIY projects ("one sense per discourse"): the most frequent paraphrase tends to be the adequate one.

Thus, we get, for example, corpus occurrences for both, compounds and their paraphrases:

- *Kupferschraube* ↔ *Schraube aus Kupfer* (material: "copper screw")
- *Befestigungsschraube* ↔ *Schraube zur Befestigung* (purpose: "fixation screw")
- *Senkkopfschraube* ↔ *Schraube mit Senkkopf* (property (or: part/whole): "countersunk screw")

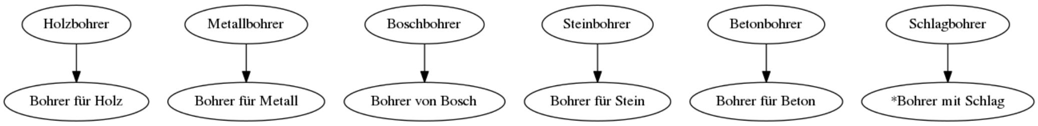

Figure 5: A subset of relations found by assigning compounds NP+PP paraphrases; arrows indicate quasi-synonymy, e.g. "*Holzbohrer* equals *Bohrer für Holz*.

Alongside the isa-relation ("*copper screw*" → "*screw*"), we can thus also extract further meaningful relations from paraphrases of compounds, cf. Figure 5. The same holds for complex NPs (*Holz der Fichte* ↔ *Holz aus Fichte* (↔ *Fichtenholz*), "spruce wood"). Obviously, some ambiguity remains: *Holzfarbe* may be paraphrased by *Farbe von Holz* ("color of wood"), as well as by *Farbe für Holz* ("color applicable to wood(en surfaces)").

3.2 Extraction and evaluation

3.2.1 Hearst-type sentences

To verify the applicability of Hearst (1992)'s approach, we implemented a German version of the classical hypernym patterns. We reproduce abstract queries (shown here in a simplified regular expression notation) in the following (where N_{sup} is the superordinate, N_{sub} the subordinate term[3]):

– N_{sub1} , N_{sub2} (und|oder) (ander.*|vergleichbar.*|sonstig.*|weiter.*) (Adj)? N_{sup}
– (Adj)? N_{sup} (,)? insbesondere (Adj)? N_{sub}

[3] The German conjunctions, adjectives and adverbs are, in sequential order "and|or", "other","comparable", "further"; "in particular"; "including"; "such as", "and|or|as well as".

– (Adj)? $\mathrm{N}_{\mathrm{sup}}$ (,)? einschließlich (Adj)? $\mathrm{N}_{\mathrm{sub}}$
– (Adi) $\mathrm{N}_{\mathrm{sup}}$wie $\mathrm{N}_{\mathrm{sub1}}$ (,)? $\mathrm{N}_{\mathrm{sub2}}$ ((’und|oder|sowie’) (Adj) $\mathrm{N}_{\mathrm{sub3}}$))*

The patterns are not mere translations of the original English patterns, but have been carefully adapted to German, including many additional constraints on the part-of-speech and lemma level to filter out wrong candidates. For example, while the EN version of pattern four ($\mathrm{N}_{\mathrm{sup}}$ such as $\mathrm{N}_{\mathrm{sub}}$) is highly effective, the German adaptation results in many wrong pairs, as in *Hubzahl wie für Baustahl* (“stroke frequency as (used) for structural steel”). Thus, we excluded e.g. results where “wie” was followed by a preposition.

Parsing is not required to identify these patterns; they can equally well be extracted from POS-tagged and lemmatized data. However, for the extraction of verbal predicates which denote class membership we have also implemented an extraction from parsed text. There, we search for the two predicates *zählen* (“be part of”) and *gehören* (“belong to”) and extract the head of their p-object as the hypernym while the head of the subject is considered to be the hyponym. We also extract predicate constructions in the form of ($\mathrm{X}_{\mathrm{sub}}$ is a $\mathrm{Y}_{\mathrm{sup}}$).

In a first evaluation, we only evaluated the POS-based nominal patterns described above. We are currently planning an evaluation of the verbal patterns[4].

We evaluated the top 200 search result pairs sorted by frequency regarding the question whether the hyponymy relation holds. This is true for 163 out of the 200 pairs, i.e. the accuracy of this technique is about 82%. Errors typically occur in pairs extracted by the fourth pattern, e.g. as in *Unterschied wie Tag und Nacht* (“difference as night and day”).

In a second version, we filtered out pairs in which none of the two nouns is a term (i.e. not in the gold standard list), sorted by frequency. We then performed a two-fold evaluation. In the first step, we looked at the validity of the hyponymy relation: do the pairs establish plausible hyponym-hypernym pairs. Out of 200 pairs, 164 were considered valid (82% accuracy). Regarding the question whether the pairs are also domain relevant, 151 out of the 164 valid pairs turned out to be domain relevant (92%).

Overall, the impression in our data is that the quality of the extracted pairs is acceptable, and many of the pairs are relevant for our domain[5].

3.2.2 Compounds

Compound analysis for taxonomic relations We split compounds using the compound splitting tool CompoST (Cap, 2014), see above. We consider the head as the superordinate, and the compounds as subtypes of their heads: *Säge (saw)* has subordinates such as *Kreissäge (buzz saw), Bandsäge (bandsaw).* The implementation is aware of complex non-heads, i.e. we check for attested morpheme combinations in our specialized corpus as well as in a large general language corpus to exclude wrong splits. For example, for *Eigenbaubandsäge* (“self-constructed bandsaw”), we first split into morphemes (Eigen| bau | band | säge) and then check for attested combinations: *Bandsäge* (valid, found), *Baubandsäge* (not found), *Eigenbau-X* (valid, found), resulting in the correct split *Eigenbau| Bandsäge*.

A script sorts all heads together with their compounds and builds a partial hierarchical structure for every head. An example hierarchy is given in Figure 4.

While these hierarchies have not yet been evaluated, their accuracy is solely dependent on the performance of the compound splitting tool. We are currently planning a comparative evaluation of several compound splitting tools to assess the quality of the compound splits. Overall, the impression when looking at a small set of these hierarchies is that they very rarely contain wrong hyponyms.

Compound analysis and paraphrases for non-taxonomic relations We acquire paraphrases for compounds of the form Noun_1+Noun_2 with nominal heads by querying Noun_2+preposition+Noun_1 or Noun_2+determiner+Noun_1 (in genitive case) in the 11M corpus. Finding nominal paraphrases for heads and non-heads of compounds helps us determine the relation between the parts of the compound. It can also help us disambiguate between possibly ambiguous relations, e.g. to decide whether a drill is

[4] The results will be available by end-November 2016.

[5] An error analysis is ongoing and will become available by end-November 2016.

Compound	Paraphrase	Relation
Steinbohrer (stone drill)	Bohrer für Stein (for)	purpose
Metallbohrer (metal drill)	Bohrer für Metall (for)	purpose
Diamantbohrer (diamond drill)	Bohrer aus Diamant (made of)	material
Heizkörperverkleidung (radiator cover)	Verkleidung vor Heizung (in front of)	location
Kellerraum (basement room)	Raum im Keller (in)	location
Schutzfolie (protection film)	Folie zum Schutz (for)	purpose
Aluprofil (aluminium profile)	Profil aus Alu (made of)	material
Pendelhubstichsäge (scroll jigsaw)	*Stichsäge ohne Pendelhub (without)	–
Wasserhaus (water house)	*Haus unter Wasser (under)	–

Table 1: Some exemplary paraphrases found in our data and the relations they indicate

(partially) made of a certain material (*Diamantbohrer – Bohrer aus Diamant*, diamond drill- drill made of diamond) or used to drill a specific material (*Steinbohrer – Bohrer für Stein*, stone drill - drill made for drilling stone). Further examples are given in Table 1. We indicate the compound, the paraphrase found in the corpus and the relation inferred by rule from the preposition. Certain prepositions, like for example *ohne* (without), are excluded as they almost never lead to relevant paraphrases.

In a precision-based evaluation, we manually evaluated the top 200 paraphrase-compound pairs, sorted by compound frequency. 157 out of 200 candidate paraphrases were valid paraphrases, resulting in 79% type accuracy. Errors are mainly due to implausible prepositions, such as *Rest im Holz* (rest in the wood) for *Holzrest* (scrap wood). Taking into account the frequencies of the paraphrases for every compound, 814 paraphrases out of 959 total paraphrase occurrences turned out to be valid paraphrases, resulting in a token accuracy of 85%.

4 Identifying events involving domain objects

4.1 Relevant phenomena

Predicate+argument-structures To find events involving the domain objects, we extract predicates and their subjects and complements as well as context information in the form of negation and adverbs.

Based on dependency output as produced by mate (Bohnet, 2010), we can extract the following categories:

- 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)

- 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)

- Predicative constructions: X is Y (Y can be adjectival or nominal):
  Bohrer ist ein Elektrowerkzeug (drill is a power tool)
  Spitze ist besonders dünn (tip is very thin)

We can also combine these extractors to search for longer patterns, including negation or adverbs.
Subj V Obj: *Holzspiralbohrer haben eine lange Zentrierspitze (wood drills have long lathe centers)*;
Subj V PP: *Beton besteht aus Zement und Wasser (concrete is made of cement and water)* ;
Subj V Obj +Negation:*Kupfer benötigt keinen schützenden Anstrich (copper requires no protective coat).*

Verb-derived items as a source of relational data Many morphologically complex words are derived from verbal (or adjectival) predicates. German is rich in noun compounds whose heads are nominalizations of verbs or adjectives (e.g. *Holzoberflächengestaltung* "design of wooden surface(s)", *Anwendbarkeit der Magnetfarbe* "applicability of magnetic colour"). Compound participles are equally productive and allow for an analysis of the underlying verbal element in terms of its predicate-argument structure (cf. *alumimiumbeschichtete Oberfläche*, "aluminium-coated surface").
Also here, the combination of compound splitting and search in syntactically annotated data provides pairs of terms and their paraphrases, where the latter make the relations explicit that exist between the items involved (see Figure 6). Alongside the above mentioned complex NPs, we also find verb+complement constructions, such as *Holzoberfläche*$_{\text{Obj}}$+ *gestalten* (to design a wooden surface), *Magnetfarbe*$_{\text{Obj}}$+*anwenden* (apply magnetic color) or *Oberfläche*$_{\text{Obj}}$+ *mit Aluminium beschichten* (coat surface with aluminium). We exploit not only verb+object pairs, but also verb+PP groups, subject+verb groups and predicative constructions. In all cases, we start from morphologically complex items and search their paraphrases. In addition, paraphrase patterns can also be exploited, in the sense of "knowledge-rich contexts" (Meyer, 2001) as models or types of events with instances which do not correspond, in the available data, to morphologically complex items: compound participles of the type *aluminiumbeschichtet* correspond to a pattern such as *X[agent] beschichtet Y[target] mit Z[coating]*, where the expressions in brackets are taken to be informally noted participant roles similar to Frame Elements of FrameNet (cf. Ruppenhofer et al. (2013)) . This pattern provides a large number of pairs of domain objects related by the ad-hoc relation "coated with", most of which are relevant for the domain and correctly recognized[6].

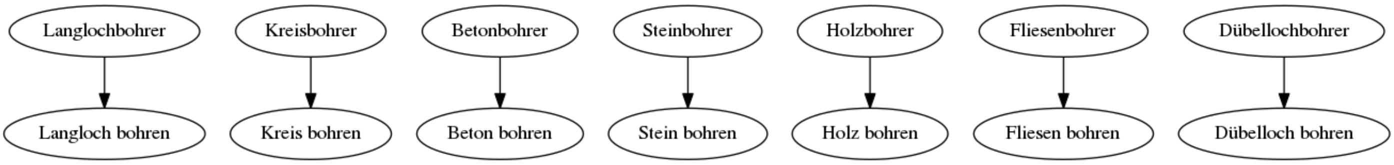

Figure 6: A subset of events found for *Bohrer* by matching compounds and their verbal paraphrases; arrows indicate a corresponding event.

4.2 Extraction and evaluation

4.2.1 Verb and object

This section describes the evaluation of verb object pairs, such as *Dübelloch bohren (drill dowel hole), Sägeblatt verwenden (use saw blade)*, or *Fliesen verlegen (lay tiles)*.

We first evaluated whether the extracted verb object pairs are syntactically valid. Thus, we manually checked the top 250 pairs ranked according to their termhood measure (in this case: domain specificity value (Ahmad et al., 1992)), only looking at pairs with a frequency > 5. The decision is made given an example sentence. Out of the 250 top pairs, 15 are syntactically invalid due to pre-processing and parsing errors. This means that 94% of the extracted pairs are syntactically plausible. Therefore, the parsing quality, although not trained on data from the DIY domain, seems well-suited as a basis to extract data.

A second evaluation looks at the question of domain relevance. Again, we analyze the 250 top ranking pairs V-O candidate pairs sorted by the ranking measure, excluding the verbs *haben, sein* and *geben* (have, be, give). The decision in this case was made between the categories "term", "no term" and "preprocessing error". 27 errors occurred (10%) due to preprocessing or parsing errors. 150 out of the 250 candidates are good terms (60%), whereas 73 bad terms (30%) are not relevant for our domain. Bad terms very often occur only because part of the subcategorization of the verb has not been covered by the extraction pattern, such as in
Werfen Sie Elektrowerkzeuge nicht in den Hausmüll ⇒*Elektrowerkzeuge*$_{OBJ}$+*werfen* $_{V}$

[6] An evaluation is ongoing. Results will be available by December.

(Do not throw power tools into the trash ⇒ throw$_V$+ power tools$_{OBJ}$)
These cases can be excluded by using longer patterns involving PPs and negation. Sometimes, the area between terms and non-terms is blurred, e.g. in *Alurohr umdrehen (turn aluminium tubes), Fliesenkleber benötigen (require tile cement), Kochfeld einbauen (assemble hob).* While these may not be top terms, they definitely are not general terms, either.

4.2.2 Verb and p-object

We performed a top 200 precision-based evaluation, assessing the verb PP pairs according to the question whether the pairs are syntactically valid. We found that 191 of 200 are syntactically plausible, resulting in an accuracy of 96%.
Most of the extracted pairs are very relevant to the domain, such as

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 are PP attachment problems, such as in *Ich suche ein Gerät mit Akkubetrieb (I'm looking for a device with battery operation) – [suchen mit Akkubetrieb].* The user generated content is also clearly visible in the extracted pairs, for example in *um einen Hammer abwerten – Ich werte um einen Hammer ab wegen der schlechten Bedienung (Giving this one hammer less due to bad usability)* or in *an die Schraube glauben – Ich glaube an die Schraube (I believe in this screw).*

4.2.3 The role of coreference resolution for the enhancement of recall in the extraction of predicate argument structures

Many times, arguments of verbs are pronominalized. In order to make use of them for relation extraction, we need to resolve them using a coreference resolver. Thus, we performed some experiments on the use of a state-of-the-coreference resolver (IMS HotCoref (Rösiger and Kuhn, 2016)) for verb object extraction. Coreference resolution in user-generated texts is considered difficult, as there is a decrease in performance of the pre-processing tools when they are used on non-standard data. We only evaluate the quality of coreference resolution indirectly, by looking at the verb object pairs extracted.
We found that, in our data, about 40% of the verb object pairs contained pronominalized objects. One assumption about using coreference resolution therefore was that we can get more candidate pairs. This is true, as the number of verb object pairs rose from 3996 to 4189 candidates (+5%). We further checked whether the newly found candidates are good candidates. We found 82% of the 193 new candidates relevant to the domain, e.g. *120er-Schleifpapier verwenden (use 120-grit sandpaper), 6-mm-Loch bohren (drill 6-mm hole).* We also found more evidence for pairs already retrieved from the version without coreference resolution, in the form of higher frequencies. We expect the assumptions proven to be true for verb object pairs to be true for other arguments as well, such as subjects or p-objects.

4.2.4 Compound analysis and verbal paraphrases

For compounds with nominalized verbs as heads, we can search for verbs and their respective object as the non-head of the compound. If we find a match, this is evidence that the compound describes an event corresponding to the verb and its object.

Compound	Paraphrase
Abflussreiniger (drain cleaner)	Abfluss reinigen (clear drain)
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)

We evaluated the 125 most frequent and the 125 least frequent compounds for which a verb+object paraphrase was found with respect to the question whether the verb object-paraphrase was valid for the given compound. The analysis of the top 125 resulted in an accuracy of 74%, for the bottom 125 the accuracy was 82%.

5 Related work

Our work applies a set of strategies that have been introduced in the literature on German user-generated and expert text. Corpus-driven ontology creation has been proposed in many papers, e.g. in Barrière (2004), Auger and Barrière (2008) or Manser (2012), to name only a few. However, to the best of our knowledge, we are not aware of any papers that test and extend these strategies on German texts.

Similar to our strategy is the approach by Gillam et al. (2005) which is also based on hybrid terminology extraction; cf. also Drouin (2003)'s approach. They apply a number of collocation and linguistic patterns to extract relations between terms from specialized English texts. Arnold and Rahm (2014) extract semantic concept relations for German terms from Wikipedia definitions. However, this approach is dependent on Wikipedia sites (i.e. expert text) and not easily applicable to user-generated text. Joslyn et al. (2008) present a distributional semantics approach, where they apply the lattice theoretical technology of Formal Concept Analysis to relations of predicates extracted from a corpus. Even though 11M words is a comparatively "large" amount of material for specialized texts, it may not necessarily be enough for a distributional approach. We also intend to be able to work on smaller corpora.

There are many papers building on the patterns described by Hearst (1992). In the approach by Snow et al. (2005), hypernym-hyponym-pairs are collected firstly by using WordNet. Then a corpus is used to find sentences in which both nouns of the pair occur. The dependency paths of the matched sentences are extracted and used as features for a classifier to determine if an unseen pair of nouns describes a taxonomic relation. Fundel et al. (2007) focus on the extraction of biomedical relations, e.g. the interaction between proteins. Dependency paths connecting the proteins of a given pair are extracted before a set of rules for filtering information is applied. This, of course, extracts relations beyond standard taxonomic ones, such as "A regulates B", but the dependency parse based approach is also applicable on the hypernym-hyponym pair detection. Maynard et al. (2009) differentiate between instance-class and subclass-superclass relations. Only persons, organizations and locations are considered as instances whereas other noun phrases are classes, extracted by patterns including "classification verbs" like *fall into, group into* or *contain* (cf. *zählen, gehören zu*, above). Zouaq et al. (2012) claim that the extraction of relations with lexico-syntactic patterns is an important basic step in structuring data that requires post-processing steps of filtering operations. Their patterns are classified into hierarchical relation patterns (also reusing Hearst Patterns) and patterns for conceptual relationships. e.g. verb (subject, object)-relations. Evaluations showed that the hierarchical patterns achieved the highest precision without post-processing of the results.

The approach described in Ritter et al. (2009) also starts with the extraction of relations using Hearst Patterns. They then filter the matches by using different methods. As applying a frequency based classifier is not sufficient, a SVM classifier is implemented to rate every extracted pair in terms of correctness. As features, a variety of frequencies is used. Finally, they develop an HMM language model to make an evaluation possible even if a certain noun does not have a match with any of the Hearst Patterns.

6 Conclusions and future work

We presented a set of techniques to acquire semantic relations between terms and showed that overall, one can achieve acceptable precision when applying standard tools to relation extraction. Future work will include more morpho-syntactic patterns to extract such relations, as well as external knowledge sources such as e.g. BabelNet. While our work focused on precision-based evaluations of highly frequent cases for the single techniques, more detailed evaluations are planned on the combination of the approaches presented here, as well as the creation of a gold standard, to also be able to assess recall.

Acknowledgments

The work reported in this paper has been carried out in the framework of the collaborative research project "Textanalyse und Ontologieaufbau" financed by the corporate research department of Robert Bosch GmbH. We gratefully acknowledge this support.

References

Khurshid Ahmad, Andrea Davies, Heather Fulford, and Margaret Rogers. 1992. What is a term? - the semi-automatic extraction of terms from text. In *Translation Studies - An Interdiscipline*, pages 267 – 278. Selected papers from the Translation Studies Congress, Vienna.

Patrick Arnold and Erhard Rahm. 2014. Extracting semantic concept relations from wikipedia. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, WIMS '14, pages 26:1–26:11, New York, NY, USA. ACM.

Alain Auger and Caroline Barrière. 2008. Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology: international journal of theoretical and applied issues in specialized communication*, 14(1):1–19.

Caroline Barrière. 2004. Building a concept hierarchy. *Terminology*, 10(2):241–263.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China*, pages 89–97. Association for Computational Linguistics.

Fabienne Cap. 2014. Morphological processing of compounds for statistical machine translation. Dissertation, Institute for Natural Language Processing (IMS), University of Stuttgart.

Béatrice Daille. 2007. Variations and application-oriented terminology engineering. pages 163 – 177.

Patrick Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium.

Gertrud Faaß and Kerstin Eckart. 2013. Sdewac a corpus of parsable sentences from the web. In *Language Processing and Knowledge in the Web: 25th International Conference, GSCL 2013, Darmstadt, Germany, September 25-27, Proceedings*, pages 61 – 68.

Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2007. Relex – relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

Lee Gillam, Mariam Tariq, and Khurshid Ahmad. 2005. Terminology and the construction of ontology. *Terminology*, 11:55–82.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cliff Joslyn, Patrick Paulson, and Karin Verspoor. 2008. Exploiting term relations for semantic hierarchy construction. In *Semantic Computing, 2008 IEEE International Conference on*, pages 42–49. IEEE.

Peter Koch and Wulf Oesterreicher. 1985. Sprache der nähe–sprache der distanz. *Romanistisches Jahrbuch*, 36(85):15–43.

Phillip Koehn and Kevin Knight. 2003. Feature-rich statistical translation of noun phrases. In *Proceedings of ACL 2003*.

Mounira Manser. 2012. État de l'art sur l'acquisition de relations sémantiques entre termes: contextualisation des relations de synonymie. In *Actes de la conférence JEP-RECITAL*, pages 163–175.

Diana Maynard, Adam Funk, and Wim Peters. 2009. Sprat: a tool for automatic semantic pattern-based ontology population. In *International conference for digital libraries and the semantic web, Trento, Italy*.

Ingrid Meyer. 2001. Extracting knowledge-rich contexts for terminography. *Recent advances in computational terminology*, 2:279.

Alan Ritter, Stephen Soderland, and Oren Etzioni. 2009. What is this, anyway: Automatic hypernym discovery. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*, pages 88–93.

Ina Rösiger and Jonas Kuhn. 2016. IMS HotCoref DE: A Data-driven Co-reference Resolver for German. In *Proceedings of LREC 2016*.

Josef Ruppenhofer, Hans C. Boas, and Collin F. Baker. 2013. The framenet approach to relating syntax and semantics. In Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, and Herbert Ernst Wiegand, editors, *Dictionaries. An international encyclopedia of lexicography*, volume Supplementary volume: Recent developments with special focus on computational lexicography, pages 1320–1329. De Gruyter.

Johannes Schäfer, Ina Rösiger, Ulrich Heid, and Michael Dorna. 2015. Evaluating noise reduction strategies for terminology extraction. In *Proceedings of TIA 2015*, Granada, Spain, November.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, pages 777 – 784.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. Smor: A german computational morphology covering derivation, composition, and inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1263 – 1266, Lisbon, Portugal.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1297–1304. MIT Press.

Amal Zouaq, Dragan Gasevic, and Marek Hatala. 2012. Linguistic patterns for information extraction in ontocmaps. In *Proceedings of the 3rd International Conference on Ontology Patterns-Volume 929*, pages 61–72. CEUR-WS. org.