Creating a gold standard corpus for terminological annotation from online forum data

Anna Hätty
Robert Bosch GmbH
Anna.Haetty@de.bosch.com

Simon Tannert
University of Stuttgart
simon.tannert@ims.uni-stuttgart.de

Ulrich Heid
University of Hildesheim
heid@uni-hildesheim.de

Abstract

We present ongoing work on a gold standard annotation of German terminology in an inhomogeneous domain. The text basis is thematically broad and contains various registers, from expert text to user-generated data taken from an online discussion forum. We identify issues related with these properties, and show our approach how to model the domain. Futhermore, we present our approach to handle multiword terms, including discontinuous ones. Finally, we evaluate the annotation quality.

1 Introduction

Terms are linguistic expressions typical of specialized domains (Kagueura and Umino (1996)). In this work, texts from the domain of do-it-yourself instructions and reports (DIY) are chosen as basis for a gold standard annotation of terminology. The DIY domain is characterized by a broad range of topics, and our text corpus in addition covers several registers. This results in the presence of term candidates with different status and poses a challenge to the annotation approach. We describe our way to model the degree of termhood and the relation of multiword terms to their variants. The model serves as a basis to define rules to limit an outgrowth of term candidates. The gold standard is intended to be a reference dataset for automatic term extraction. Such a system has to cope with heterogeneous domains, with morphologically related term variants as well as with variants emerging from the different styles present in the text corpus.

In the following, our domain and annotation approach are positioned on the map of existing term annotation work. In section 3, we describe how the text basis is chosen to ensure that it is representative of the DIY domain. In section 4, we describe the annotation procedure and address the challenges that arise from the selected domain and registers. Finally, our annotation is evaluated, and we interpret systematic divergences between annotators. We conclude in section 5.

2 Related Work

Existing Benchmark Datasets for Term Extraction There exists a range of terminology benchmark datasets which vary in the specificity of their topic, their definition of termhood and writing styles. Well-known datasets are the **Genia** (Kim et al. (2003)) and the **CRAFT corpus** (Bada et al. (2012)) with term annotations in the biomedical domain. Genia contains 2000 MEDLINE abstracts with almost 100,000 annotations by two domain experts. CRAFT consists of 67 biomedical journal articles of various biological domains (plus unpublished articles) with more than 100,000 concept annotations. **ACL RD-TEC** (Handschuh and QasemiZadeh (2014)) is a gold standard in the domain of computational linguistics. It consists of 10,922 ACL conference papers published between 1965 and 2006. From those more than 83, 000 term candidates have been extracted and evaluated; 22,000 candidates are annotated as valid and 61,000 as invalid terms by one annotator. An extension is **ACL RD-TEC 2.0** (QasemiZadeh and Schumann (2016)), a further annotation of 300 ACL abstracts with a broad subclassification of the terms.

corpora	ACL 1.0	ACL 2.0	B/C	Bitter	TTC	Genia	Craft	our approach
breadth	**	**	**/*	**	**	*	**	***
registers	*	*	*	*	**	*	*	***
token-based	-	+	+	+	-	+	+	+
guidelines	broad	broad	mid/strict	broad	mid	strict	strict	mid

Table 1: Comparison of terminology gold standards

Bernier-Colborne and Drouin (2014) (B/C) analysed three textbooks on automotive engineering. In addition to the annotation, they assign attributes to the terms (e.g. for acronyms or multiwords) and mark orthographic variants. Other reference sets consist of bilingual term lists to evaluate machine translation. In the **TTC project** (Loginova et al. (2012)), a list of term candidates is generated with a term extraction tool and then further evaluated by experts. In the **BitterCorpus** (Arcan et al. (2014)), terms are annotated in texts from KDE and GNOME documentation corpora. In the following, we compare the reference datasets wrt. the size of their domain, the registers represented and the underlying annotation approach (see also Table 1).

Domain. The reference datasets differ wrt. the breadth of the topics covered. Genia's domain is very narrow, it is specialized to biological reactions concerning transcription factors in human blood cells. The texts are crawled on the basis of three seed terms. With Bernier-Colborne and Drouin, the topic is automotive engineering as presented in three textbooks for lay people. For CRAFT and ACL RD-TEC, journal and conference articles have been taken from a wide range of subtopics in their respective domains, and different research areas of the domains are included in the text basis. The same holds for the BitterCorpus: In the GNOME and KDE manuals, a range of topics, such as the user interface, settings, the internet connection or information about hardware are addressed. All these corpora have clearly defined content since the extraction basis is hand-selected. This does not hold for the TTC texts, which are retrieved by a thematic web crawler; unexpected text can thus occur in the corpus. The topics of our own data are even more open: The DIY domain is broad in itself, and as the texts come from different sources, the variety of topics even increases. Several slightly off-topic texts are part of the text basis.

Register. Most of the gold standard corpora are homogeneous wrt. register. They either consist of scientific articles (Genia, CRAFT, ACL RD-TEC 1.0 and 2.0) or of instruction texts: The three expert-to-lay textbooks for automotive engineering might differ slightly from author to author, but nevertheless have the explanatory style of textbooks. Finally, the KDE and the GNOME documentation follow the style of online manuals. Different registers only occur in the crawled text of TTC. In our work, we deliberately chose texts from different registers and sampled the text basis in a way that expert writing and user generated content (= UGC) are represented both (60:40%).

Annotation Approach. The definition of termhood is widely divergent across the different gold standards. In Genia and CRAFT, the annotation is very strict, as specific syntactic patterns and semantic contraints are given. Both the work by Bernier-Colborne and Drouin (2014) and the TTC terms have a more liberal annotation scheme, partly following the rules proposed by L'Homme (2004). Bernier-Colborne and Drouin (2014) limit the annotation semantically to items denoting components of cars and for TTC, term candidates were preselected by a term extraction tool. For the ACL RD-TEC gold standards and the BitterCorpus, the definition of termhood is particularly liberal, as termhood is rather loosely defined. They mainly rely on the association an annotator has with respect to a term or to a domain (e.g. by structuring terms in a mindmap) and provide theoretical background about terminology.

For our work, we aim at a compromise between generality of annotation and restriction of outgrowths. Because of the breadth and the stylistic variability of the DIY text basis, we do not set strict

corpora	total	used	corpora	total	used	corpora	total	used
wiki	4.31 * 10^5	30,915	FAQs	4,805	347	project	2.16 * 10^6	2,701
expert projects	55,430	3,971	encyclopedia	6,059	449	forum	2.34 * 10^7	29,293
marketing texts	35,452	2,540	book	54,005	3,868			
tips and tricks	12,711	904	tool manuals	69,831	5,012			

Table 2: Distribution of tokens by subcorpus: expert (two left-most) and user texts (right)

rules for the annotation, e.g. by limiting the syntactic or semantic shape of terms by predefined POS-patterns or predefined ontology elements onto which the terms would have to be mapped. However, we give positive and negative examples, and guiding rules elaborated after extensive discussion about the relation of DIY terms to their domain.

3 Corpus and Domain: from User-Generated to Standard Text

We use a corpus of German texts of the DIY domain, which is thematically about non-professional builds and repairs at home. There are different text sources available, containing texts produced by domain experts as well as by interested lay users. The latter mainly consists of forum posts collected from several online DIY-forums, e.g. from project descriptions or inquiries for instructions. Experts texts include an online encyclopedia and a wiki for DIY work, tools and techniques. The corpus used for the work described here contains ca. 11 M words in total, with 20% expert text vs. 80% user-generated data.

For the manually annotated part, we aim at a balanced extraction of text data from all the different sources. Thematically, we only excluded gardening activities, which we do not see as a part of the DIY domain. The corpus is balanced to include 40% user texts and 60% expert texts. In total, 80,000 tokens are extracted. Since we annotate terms in context (token-based), complete sentences are extracted. We thus sample subcorpora proportionally to their orginal size, to reach a total of 48,000 tokens of expert text plus 32,000 tokens of UGC (see Table 2). All sentences are shuffled.

4 Annotation

4.1 Procedure and Design of Annotations

General Procedure The annotation guidelines were created in discussion rounds with 6 to 7 participants who have experience in terminology extraction. All are semi-experts of the domain, because they have been dealing with terminology extraction from the DIY domain for more than one year. The guidelines were influenced by terminology theory, peculiarities observed when analysing the text data and practical issues, to ensure a consistent annotation. The actual annotation is being produced by three (of the above) annotators; at the time of writing, two annotators have finished half of the corpus, i.e. 40,000 tokens are annotated.

Annotation Tool We use **WebAnno** (Yimam et al. (2013), de Castilho et al. (2016)) as an annotation tool, a multi-user tool with a several annotation modes and a visualisation of the annotation. In our case, possible annotations are **spans** (for single- and multiword terms) and **relations** (used here to link separated parts of a term). For the spans, several **values** can be chosen: *domain, domain-zusatz* and *ad-hoc*. While most terms are annotated with *domain*, we use *ad-hoc* for user-coined terms, and *domain-zusatz* (= domain-additional element) for elements that are themselves not terms, but are parts of multiword expressions, e.g. the adverb *freihand* in *freihand sägen*.

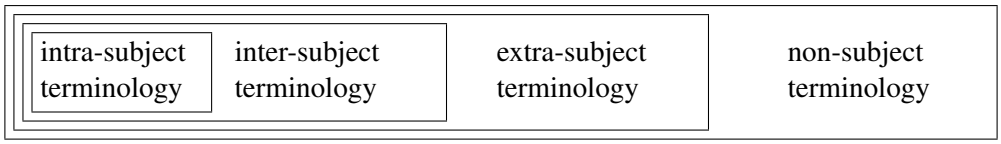

Figure 1: Tiers of terminology (Roelcke (1999)) [our translation]

4.2 Tiers of Terminology and Consequences for the Annotation Approach

The annotation of benchmark sets for terminology is typically implemented as a binary decision. However, it is widely acknowledged that the terminology of a domain is a rather inhomogeneous set. It can be divided into several tiers, e.g. with a distinction between terms which only occur in the very specific vocabulary of a small domain, as opposed to terms which occur with an extended or specialized meaning in one domain but also in other domains or in general language (e.g. Trimble (1985), Beck et al. (2002)). The model by Roelcke (1999) consists of four layers (Figure 1), where the most restrictive one is the **intra-subject terminology** which is specific to the domain. The **inter-subject terminology** is used in the respective domain, but also in others. The **extra-subject terminology** is terminology which does not belong to the domain but is used within it (we call such terms *borrowed terms*). The last group, the **non-subject terminology**, consists of all items used across almost all specific domains. Tutin (2007) calls this 'transdisciplinary vocbulary': it includes the domain-unspecific language of scientific writing (e.g. *evaluation, estimation, observation*) and non-specialized abstract vocabulary (e.g. *to present a problem, to result in*).

Our annotation approach is liberal and our notion of termhood comprises the first three layers of Roelcke's model. As a consequence, often no clear borderline between the DIY domain and other domains can be drawn; following the example of the TaaS project (www.taas-project.eu), we therefore provide the annotation with confidence scores about how many of the annotators agreed on the annotated element to be a term and distinguish between a strict (three annotators agree) and a lenient (two of three annotators agree) annotation.

4.3 Breadth of the Domain: Terminological Richness in the DIY-Domain

The DIY domain is influenced to a high degree by other domains. There is a quite obvious core set of terms which are prototypical (e.g., *drill, fretsaw, circular saw bench, ..*). In addition, there are many terms borrowed from other domains, e.g. from material science or construction techniques. In our annotation, we distinguish between **terminology borrowed from other domains** and **terminology from neighbouring domains**. While texts with intra-subject or inter-subject terms tend to centrally belong to the DIY domain (and describe what we consider to be "typical" DIY activities), borrowing takes place from related domains knowledge about which is necessary for efficient communication in the DIY domain, such as some fields of physics, of material science, construction techniques, etc. We consider fields as neighbouring domains which are carried out professionally, such as sanitary, electrical or heating engineering. Sentences belonging to texts describing work of this kind are disregarded in our annotation.

4.4 Registers: User Language and Jargon

Apart from the broad domain, the wide range of registers is a challenge for annotation. In the user-generated texts, misspellings and user-coined terms (e.g. *Selberbauer, reinhämmern, Filterabrüttlung, mit Hobelmesser "abgemessert"*) have to be addressed. We mark them with the special label *ad-hoc*, to show their terminological relevance but to distinguish them from accepted terms.

The way in which DIY-forum users talk about tools and materials shows their high degree of specialization, even in texts that exhibit signs of conceptual orality (in the sense of Koch and Oesterreicher (1985)). In the 40.0000 words, we identified 71 references of tools in which a highly specialized DIY knowledge is presupposed:

From the standard (expert) text in the domain, we observe that the official denomination of power

tools mostly follows a rigid pattern. The names are composed of [BRAND][TYPE][MAIN DESIGNATION][SECONDARY DESIGNATION], for example *Metabo Kappsäge KGS 216 M* or *Bosch Tischkreissäge PTS 10 T*. An intuitive way of abbreviating those denominations would be by the type; instead we find highly specific references, close to in-group jargon:

- 16 times the tool was only referenced by its brand name (e.g. *meine Makita, Metabo, ...*);
- 24 times by its main designation (*IXO, PBS, ...*);
- three times by its secondary designation (*0633 ohne Motor, 900er*);
- and 28 times by a combination of main and secondary designation - of different granularity and written in different forms (*GKS 68 BC, PCM8S, ...*).

This special term use increases the number of term types and poses a challenge for automatic term extraction, as well as for coreference resolution in that domain. Furthermore, this way of referencing supports the claim that embedded terms need to be addressed in the manual annotation. Whether a term extraction tool which is sensitive to embedded terms can also identify this kind of references, is still an open question. There are less regular references as well, e.g. abreviations by material (*ODF* instead of *ODF-Platte*), missing size units (*35er Scharnier*), or only sizes are mentioned (*K60-K220* instead of *Schleifpapier der Körnungen K60, K80, .., K220*)). Other special cases are jargon-like abbreviations (*TKS = Tischkreissäge*, *OF = Oberfräse*, *HKS = Handkreissäge*).

Another characteristic of user texts is the almost infinite number of domains from where terms can be borrowed: when being creative, everything can be used to do handicrafts with, everything can be (mis)used as a tool or material (*Frühstücksbrett in Fliesenoptik; Geschenkboxen aus Käseschachteln, gedrechselte Kirschen*). Items from these other domains fill areas in DIY which are prototypical, e.g. DIY project names, materials and tools. This makes it harder to decide whether these items are terms. That topics are spread more widely can be shown by the number of sentences annotated in the 40.000 corpus: In the user-generated content (UGC) part, 45.36% of the sentences are annotated, in the expert texts 66.21%. Furthermore, the density of term annotation is higher for the expert texts: in the UGC texts, 9.15% of the tokens are annotated, in the expert texts 17.08%.

4.5 Annotation Approach: Multiword Terms and Term Variants

A special focus of the annotation is on multiword terms (MWTs). We aim to preserve as much of the terminological content in the data as possible. By allowing to annotate discontinuous multiword terms, we enrich the term base.
Besides annotating adjacent MWTs, we also capture MWTs interrupted by terminologically irrelevant material. In *scharfes und gefährliches Messer* (sharp and dangerous knife) *und gefährliches* will not be annotated, while *scharfes Messer* is considered as a term. This annotation is realized by linking together the separate parts of the MWT. A similar case are MWTs which are interrupted by additional terminological material, e.g. *schwebender (schwibbender) Bogen*, from where two terms can be created by linking: *schwebender Bogen* and *schwibbender Bogen*.
Contrary, e.g. to TTC, we annotate all valid embedded terms. For example, for *freihand gebrochene gerade Kante*, the whole term, *gerade Kante* and *Kante* are annotated.

As we aim at covering all possibly terminologically relevant material, we do not a priori set restrictions as to the length or POS pattern of term candidates. Anyway, collocational verb-noun pairs (*Holz fräsen, mit Nägeln verbinden*) are not annotated as multiword terms. We aim at distinguishing them from terms. However, this annotation decision leads to an inconsistency at the theoretical level: If the verb-noun pair occurs in its nominalized form (*Nagelverbindung*). As a consequence, we annotate the noun compound form and have this inconsistency; to attenuate this conflict, we also allow idiomatic verb-noun combinations to be annotated. For example in *auf Gehrung sägen*, *auf Gehrung* is annotated as **domain-zusatz** ('domain additional element') to *sägen*.
Our annotation keeps track of the variety and complexity of syntactic structures in which terms can appear in texts, including non-adjacent parts of multiword expressions.

5 Evaluation

5.1 Inter-Annotator Agreement

Fleiss' kappa (Fleiss et al. (1971)) is used to calculate the inter-annotator agreement. In our annotation, multiword terms, parts of terms and different annotation labels have to be considered. In total, 2514 single-word terms (SWTs) and 511 MWTs are annotated by one annotator, 4269 SWTs and 1353 MWTs by the other one. An item can have multiple labels. Thus, we introduce an IOB format for the terms (term-internal, out-of-domain, beginning of a (multiword) term) and consider the annotation to have 9 labels: IOB * labels *domain, ad-hoc, domain-zusatz*. Fleiss' kappa is calculated for every label and the result is averaged. We achieve an interannotator agreement of 0.81 which is a substantial agreement according to Landis and Koch (1977).

5.2 Error Analysis: Consistent Differences in MWTs Annotation

Despite our strategy to encourage the annotation of MWTs as well as of their embedded terms, we still find consistent differences in this regard. Two kinds of structural inconsistencies are prevalent:

Adj N In 151 out of 455 adjective-noun sequences annotated in total (by either of the annotators), one annotator annotated the whole phrase while the other one annotated only the noun. When analysing the relevant phrases, it is striking that in these cases the adjectives are evaluative (*handliche Fräse*), uninformative (*gängiger Handhobel*), underspecified dimension adjectives (*präziser Schnitt*) or related to the given situation (*vordere Schleifplatte*).

N Prep N In 17 out of 86 cases a noun-preposition-noun phrase is annotated as one stretch by one annotator while the other annotator distinguishes between two single word terms. This set consists of nominalized verb-object pairs (*Schleifen von Kanten*), positional descriptions (*Querlöchern in Holzwerkstoffen*) and purpose constructions (*Sägeblätter für Porenbeton*).

We could refine the guidelines down to individual syntactico-semantic patterns (e.g. positional vs. purpose N Prep N groups), but this would not allow us to take the linguistic creativity of the forum authors into account. Similarly, the vagueness of underspecified dimension adjectives seems rather to be the typical property of the style of our texts. As a consequence, the terms extracted from the forum data can at best be partly organized in ontologies.

6 Conclusion

We presented work towards a benchmark dataset for terminology extraction in the DIY domain. Challenges for annotation are the breadth of the domain and the register variety in our corpus. The corpus is characterized by its heterogeneity, as illustrated by a comparison of expert and user-generated text: User-generated text both has a lower density of terms than expert text (expectably) and jargon-like intra-community terminology. The domain as well as the text characteristics of UGC require specific provisions for the different tiers of terminology they contain (e.g. borrowed terms, neighbouring domains).
Our annotation approach is liberal, yet based on precise guidelines where this is realistic. We pay special attention to the annotation of multiword terms including discontinuous ones. We achieve a substantial inter-annotator agreement for the annotation.
At the time of writing, 40,000 tokens are annotated by two annotators. The dataset will be extended to 80,000 tokens and 3 annotators. We are negotiating the right to publish the annotated dataset.
Future work will include the test of term extraction tools against the dataset, possibly an additional annotation of verb+object pairs, as well as an (automatic) annotation of all sentences with markers for conceptual orality (Koch/Oesterreicher). This may provide more evidence about the relationship between register, style and terminology in forum data.

References

Arcan, M., M. Turchi, S. Tonelli, and P. Buitelaar (2014). Enhancing statistical machine tranlsation with bilingual terminology in a cat environment. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, pp. 54–64.

Bada, M., M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W. A. B. Jr., K. B. Cohen, K. Verspoor, J. A. Blake, and L. E. Hunter (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics 13*, 161.

Beck, I. L., M. G. McKeown, and L. Kucan (2002). *Bringing words to life*. New York, NY: The Guilford Press.

Bernier-Colborne, G. and P. Drouin (2014). Creating a test corpus for term extractors through term annotation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 20*(1), 50–73.

de Castilho, E., . R., Mújdricza-Maydt, S. Yimam, S. Hartmann, I. Gurevych, A. Frank, and C. Biemann (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the LT4DH workshop at COLING 2016*, Osaka, Japan. Association for Computational Linguistics.

Fleiss, J. et al. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin 76*(5), 378–382.

Handschuh, S. and B. QasemiZadeh (2014). The acl rd-tec: a dataset for benchmarking terminology extraction and classification in computational linguistics. In *COLING 2014: 4th International Workshop on Computational Terminology*.

Kagueura, K. and B. Umino (1996). Methods of automatic term recognition: A review. *Terminology* (3(2)), 259–289.

Kim, J.-D., T. Ohta, Y. Tateisi, and J. Tsujii (2003). Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics 19*(1), 180–182.

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

Landis, J. R. and G. G. Koch (1977). The measurement of observer agreement for categorical data. *Biometrics 33*(1).

L'Homme, M.-C. (2004). *La terminologie : principes et techniques*. Les Presses de l'Université de Montréal.

Loginova, E., A. Gojun, H. Blancafort, M. Guégan, T. Gornostay, and U. Heid (2012). Reference lists for the evaluation of term extraction tools. In *Proceedings of the 10th International Congress on Terminology and Knowledge Engineering (TKE)*, Madrid, Spain.

QasemiZadeh, B. and A.-K. Schumann (2016). The acl rd-tec 2.0: A language resource for evaluating term extraction and entity recognition methods. In *LREC*.

Roelcke, T. (1999). *Fachsprachen*. Grundlagen der Germanistik. Erich Schmidt Verlag.

Trimble, L. (1985). *English for Science and Technology: A Discourse Approach*. Cambridge: Cambridge University Press.

Tutin, A. (2007). Traitement sémantique par analyse distributionelle des noms transdisciplinaires des écrits scientifiques. *In Actes de TALN*.

Yimam, S. M., I. Gurevych, R. Eckart de Castilho, and C. Biemann (2013, August). Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Sofia, Bulgaria, pp. 1–6. Association for Computational Linguistics.