Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020), pages 80–84
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Computational Aspects of Frame-Based Meaning Representation in Terminology

Laura Giacomini, Johannes Schäfer
Institute for Information Science and Natural Language Processing, University of Hildesheim
Universitätsplatz 1, 31141 Hildesheim (Germany)
laura.giacomini@uni-hildesheim.de, johannes.schaefer@uni-hildesheim.de

Abstract

Our contribution is part of a wider research project on term variation in German and concentrates on the computational aspects of a frame-based model for term meaning representation in the technical field. We focus on the role of frames (in the sense of Frame-Based Terminology) as the semantic interface between concepts covered by a domain ontology and domain-specific terminology. In particular, we describe methods for performing frame-based corpus annotation and frame-based term extraction. The aim of the contribution is to discuss the capacity of the model to automatically acquire semantic knowledge suitable for terminographic information tools such as specialised dictionaries, and its applicability to further specialised languages.

Keywords: frame-based terminology, term extraction, technical terminology

1. Introduction

In the context of a larger study on variation in technical terminology carried out at the Institute for Information Science and Natural Language Processing of Hildesheim University, we have devised and implemented a method for ontology- and frame-based term variation modeling for texts concerning technical products. In this paper, we will concentrate both on already performed tests and on ongoing work. Our aim is to introduce our frame-based model and its advantages for representation of term meaning in lexicographic and terminographic resources, providing details on our method for frame-based corpus annotation, ranging from corpus preprocessing to semantic labeling.

Examples cited in this paper come from a 5.2-million-word corpus of specialised German texts concerning thermal insulation products and specifically built for this project.

2. Synonymous Term Variation

The relatively low degree of standardization of many technical subfields is one of the main reasons for the thriving of terminological variation in technical language. Synonymy, in particular, appears to be a pervasive phenomenon that is in strong contradiction with the traditional Wüsterian conception of terminology (Wüster, 1974). In particular, texts by the same source (or even the same text) often contain multiple (near) synonymous variants. These variants are sometimes characterized by the coexistence of morphologically divergent technical terms (e.g. *Hard Disk* vs. *Festplatte*, *technische Hydromechanik* vs. *Hydraulik*, *dämmen* vs. *isolieren*) but, more often, they consist of clusters of single word and multiword terms displaying morphological similarity (Giacomini, 2017). This is given in (1):

*(1) Holzweichfaserdämmplatte,
Weichholzfaserdämmplatte,
Holzfaserdämmplatte,
Holzfaserplatte zur Dämmung von...,
Platte aus Holzfasern zur Dämmung von...*

Morphological similarity is here referred to variants sharing lexical morphemes. The relationship between members of a variant cluster (such as the one in (1)) can normally be described in terms of the syntactic rules proper of a language, but their actual presence in texts may escape predictability and be motivated by contingent factors which are cognitive or discursive in nature (Freixa, 2006). We have developed a method for semi-automatically detecting variation in technical texts by relying, on the one hand, on the morphological similarity of variants and, on the other hand, on a frame-based approach to terminology (Faber, 2012/ 2015), according to which a cluster of synonymous variants takes on the same semantic role (or combination of semantic roles) within a specific conceptual scenario (frame).

3. The Frame-Based and the Ontological Description Layer

Our frame-based approach to terminology presupposes the description of the frame that is most apt to identify the topics dealt with by specialized texts contained in a corpus. It also presupposes that the technical products we aim to cover are similar in nature and function. A frame is a cognitive structure describing a situation made up of a set of specific semantic roles (frame elements, in the following named FEs) played by terms used in that situation (cf. Frame Semantics, Fillmore and Baker 2010).

On the one hand, we take into account investigations showing a comparable approach (Corcoglioniti et al., 2016, Gómez-Moreno and Castro, 2017, Anić and Zuvela, 2017 among others). On the other hand, we also look at studies concerning the automation of frame-based semantic analysis for the German general language (especially Burchardt et al., 2009), as well as studies on the application of a frame-based approach to specific semantic aspects (e.g. sentiment analysis, for instance in Reforgiato Recupero, 2015).

We specify frames with reference to a previously defined domain ontology. For the field of thermal insulation products, an OWL-specified ontology has been built based on existing resources such as the upper ontologies SUMO and DOLCE, wordnets, technical dictionaries, and specialized literature. Three ontological macroclasses, MATERIAL, FORM, and FUNCTION, are first of all identified: they include several classes of ontological entities involved in the extralinguistic reality of insulation products (Giacomini, 2017/ 2019). Among suitable frames for the description of thermal insulation products (Giacomini, 2020), we concentrate on the frame FUNCTIONALITY and manually create an initial set of core frame elements, for instance the MATERIAL of which a

product is made, the DELIVERY FORM in which a product is sold, the TECHNIQUE by means of which a product is applied, or the PROPERTY of a product.
We identified the following frame elements by analyzing corpus texts and automatically extracted candidates:

MATERIAL, MATERIAL CLASS, MATERIAL ORIGIN, MATERIAL PRODUCTION TECHNIQUE, PROPERTY, DELIVERY FORM, PACKAGING, MANUFACTURING FEATURE, TARGET, TARGET MATERIAL, COMPLEMENT, APPLICATION TECHNIQUE, TOOL, USER, PROJECT, SYSTEM, GOAL, RESULT, PRODUCT.

Each FE signals the semantic role played by a term (e.g. *Platte* (board) corresponds to the FE FORM) or part of a term (e.g. *Matte* (batt) in the compound *Steinwollematte* (stone wool batt) also corresponds to the FE FORM), and thus enables us to recognize this role across different terms, especially if they are morphologically similar. In the following example, an excerpt from a variant cluster of German terms for *extruded polystyrene insulation board* is manually annotated with POS and FE labels (e.g. N: FORM):

Platte aus extrudiertem Polystyrol :
N:FORM aus V:MAT_TECH N:MAT

Dämmplatte aus extrudiertem Polystyrol :
(V:GOAL N:FORM) aus V:MAT_TECH N:MAT

Polystyrol-Extruderschaum-Dämmplatte :
N:MAT - (V:MAT_TECH N:MAT_CLASS) - (V:GOAL N:FORM)

XPS-Platte :
(V:MAT_TECH N:MAT - N:MAT_CLASS)- N:FORM

Any term in the cluster includes the following, minimal FE combination:

MATERIAL (MAT), DELIVERY FORM (FORM), MATERIAL PRODUCTION TECHNIQUE (MAT_TECH),

whereas the frame element MATERIAL CLASS (MAT_CLASS) may additionally appear in some cases as a further specification of MATERIAL.

4. Creating a String-Based Seed Lexicon

The frame-based tagset used in our study is made up of the core frame elements found for the frame FUNCTIONALITY. In order to perform initial annotation, a number of terminological strings derived from extracted terms needs to be attributed to the frame-based tags. This leads to a seed lexicon of string-tag associations. The strings can either be full words, roots or stems depending on factors such as inflectional and derivational properties of the terms to which they belong, or their occurrence within compounds (all different cases are collected and described in a guideline).
It needs to be pointed out that a preliminary experiment of compound splitting using COMPOST (Cap, 2014) had failed to return sufficiently robust results for the German language. Moreover, the choice of employing different types of strings can be generally explained with the morphological orientation of our approach. Some string examples will be now mentioned together with the corresponding FE tag:

MATERIAL: baumwoll, glas, holz, cellulose,...
MATERIAL ORIGIN: natur, pflanz, herkunft,...
MATERIAL PRODUCTION TECHNIQUE: bläh, back,...
PROPERTY: beständig, brenn, dicht, fein,...
APPLICATION TECHNIQUE: blas, klemm, verschraub,...

For the sake of avoiding multiple and, above all, incorrect annotation, we sometimes allow for overstemming and understemming (e.g. we include all these strings: *pore, porig, porös,* and *dämm, dämmung, dämmen*). Generally speaking, priority is given to the recognition of small groups of semantically homogenous words, which is particularly important in the case of the verb *dämmen* (to insulate) and its derivatives: *dämmen*, for instance, refers to the FE GOAL, the nominalization *Dämmung* (insulation) can either refer to a GOAL, a RESULT, or a PRODUCT.

5. Semantic Annotation and Variant Extraction

The collection of technical texts is first tokenized and annotated with part-of-speech tags and lemmata using the RFTagger (Schmid and Laws, 2008). An automatic correction step is applied to make a best guess for those word forms that are unknown to the tagger lexicon. For efficient querying, the annotated corpus is then encoded for the IMS Open Corpus Workbench (CWB) (Evert and Hardie, 2011).
We then annotate these texts using the abovementioned frame elements (Section 5.1) and extract terms and variants from the encoded corpus (Section 5.2).

5.1 Semantic Annotation Employing Frame Elements

We automatically annotate tokens with the frame-based tags if they contain any of the predefined strings from the seed lexicon. Here, we exclude one frequent special case and decide not to annotate PROPERTY whenever a match of the string *offen* (open) in words containing *stoffen* (materials) is given, since this would cut the word stem. It should be noted that our string-based technique might produce other linguistically incorrect annotations, however we accept this noise for the sake of finding a higher number of potential terms in a liberal approach aiming for high recall. Tokens containing strings which are attributed to multiple frame element tags, for example the string *dämmung*, are annotated with this ambiguity, i.e. in this example GOAL/RESULT. In cases where multiple strings are matched in a single token and thus multiple frame element tags have been annotated, a special treatment to check for recursive matches is applied.
An overlapping of seed strings does not occur since they have been chosen in such a way as to exclude this. However, embedding is allowed, for example, the word *Wärmeleitfähigkeit* (thermal conductivity) contains the four PROPERTY strings *wärme*, *leitfähig*, *leit* and *fähig*. In this case, since *leit* and *fähig* are embedded in the string *leitfähig*, we only annotate *wärme* and *leitfähig* as primary

annotation and *wärme*, *leit* and *fähig* as alternative (or embedded) annotation.

In the annotation of embedded FEs, we exclude morphologically incorrect cases of string matching. First, we do not consider the string *latte* as being embedded in the string *platte*. Second, we do not consider the strings *zell* or *lose* as being embedded in the string *zellulose*. In both cases the shorter, embedded string is not annotated if the longer one is also matched. In general, our string comparison is not case-sensitive, except for strings which are specifically in upper case, for example, abbreviations such as *PUR* (Polyurethane). Finally, the annotation of matched FE tags is also encoded into the CWB corpus.

5.2 Frame-Based Extraction of Terms and Variants

We first use the IMS Open Corpus Workbench and adapt the terminology extraction approach presented in Schäfer et al. (2015) to our purposes, obtaining a list of nouns and nominal multiword candidate terms ranked according to termhood measures (for details about the termhood measures, cf. Giacomini, 2020). Category metadata are included in the output, listing for each candidate term lemma the different associated word forms, its part-of-speech annotation, and example sentences from the corpus. Term candidates and concepts from the domain ontology are employed to define a relevant frame-based tagset (cf. Section 3). This tagset, in turn, is used to semantically annotate the corpus. We then extract terms and variants using our annotation of frame element tags. In a first step, we consider all tokens which are annotated with multiple frame element tags, typically compounds. We filter these compounds by only selecting tokens with a maximum of five frame element tags, since we observed that tokens with more tags are mostly unwanted, probably results of erroneous spelling or tokenization.

Word forms are then grouped according to their frame element tags. Here we consider both primary and alternative frame element tag annotations. In a second step, we extract multiword variants for each of these compounds as follows. Initially we consider the frame element tags of a compound as a set, and compute all possible variant shapes as parts of the partition (without the original set) of this set.

For example, the compound *Vakuumisolationspaneel* (vacuum insulated panel) with the three contained strings *vakuum*, *isolation* and *paneel*, in set form: s={vakuum, isolation, paneel}, has the four different variant shapes:

s_v1 = {{vakuum}, {isolation}, {paneel}},
s_v2 = {{vakuum, isolation}, {paneel}},
s_v3 = {{vakuum}, {isolation, paneel}} and
s_v4 = {{vakuum, paneel}, {isolation}}.

Here, every set in each variant shape corresponds to a separate word, e.g. for s_v3 we would search for variants of the compound ‘s’ consisting of two words, one containing a string annotated with {vakuum} and a second one containing strings annotated with {isolation, paneel}. Furthermore, we consider every possible order of these words and consequently search for all permutations of each variant shape set. For instance, for s_v3 we take both {{vakuum}, {isolation, paneel}} and {{isolation, paneel}, {vakuum}} into consideration. However, we constrain our search to variants for which all FE-tagged words are found in a single sentence.

We group all found variants by their ordered variant shape and extract for each match the corresponding word forms and part-of-speech tags. To detect further variants when computing variant shapes, we also leave single strings associated to a certain frame element tag. For example, given the abovementioned set ‘s’, we also search for any other word annotated as FORM together with {vakuum, isolation} in a sentence, thus looking for the more general pattern {{vakuum, isolation}, {FORM}}. This is a more liberal method which produces more errors and less relevant terms, and which has therefore been employed as a secondary option.

By automatically applying ontological restrictions to FE combinations and syntactic restrictions to multiword terms, we are also able to identify previously unknown string constellations. Also extracted variants in which a component (head or non-head) is expanded, e.g. *Dachdämmung* - *Steildachdämmung* (roof insulation - pitched roof insulation) are particularly interesting, since they can potentially reveal new words which might be exploited for extending the domain ontology. We plan to release the data in 2020.

6. Statistics

In this section, we provide the results of the semantic annotation and term extraction on our 5.2-million-word corpus.

6.1 Statistics on Semantic Annotation

In total, 869,158 tokens in our corpus were matched with the defined seed strings and automatically annotated with frame element tags. Out of these, 162,462 also have an alternative annotation. Table 1 shows the distribution of the different tags in the corpus by their frequencies. Here we count occurrences in the primary and alternative annotation.

Frame Element Tag	Frequency
PROPERTY	273,129
TARGET	253,891
MATERIAL	151,924
RESULT	129,165
DELIVERY FORM	88,528
GOAL	86,774
APPLICATION TECHNIQUE	61,499
PROJECT	60,456
TARGET MATERIAL	35,322
PRODUCT	33,478
SYSTEM	28,691
MATERIAL ORIGIN	27,836
MATERIAL CLASS	14,724
MATERIAL PROD. TECHNIQUE	13,144
USER	9,917
PACKAGING	9,669
MANUFACTURING FEATURE	6,579
COMPLEMENT	4,509
TOOL	3,659

Table 1: Annotated frame elements

Figures indicated in the table correspond to the expected performance of the different frame elements: PROPERTY and TARGET are, together with MATERIAL, the conceptually most important elements of the frame, and identify the largest sets of strings in the seed lexicon. PROPERTY, in particular, comprehensively refers to chemical and physical properties of insulation products and insulation materials, but also to physical quantities. Semantic content related to insulation materials, other than in the case of PROPERTY, has been distributed across several frame elements (MATERIAL, MATERIAL CLASS, MATERIAL ORIGIN, MATERIAL PRODUCTION TECHNIQUE), which explains the lower number of tags which have been attributed e.g. to MATERIAL alone.Since we focus during extraction on compounds with multiple frame element tags, we analyze the number of tags for each annotated token. Most annotated tokens only match one of our frame element tag strings, precisely 615,171 out of the 869,158, which is approximately 71%. With an increasing number of tags per token, the frequency decreases.

6.2 Statistics on Term Extraction

Our approach to term and variant extraction uses the annotation of the predefined frame element tags with strings as previously described. As a result, we extract combinations of these annotated tags in single word terms and multiword terms at sentence level. Our 5.2-million-word corpus contains 3,124 unique word-level FE combinations (with a frequency of at least 5 to avoid excessive fragmentation).

Each base term lists any of the possible variants with their corresponding word forms if they were found at least five times in the corpus. Table 2 shows the distribution in numbers of variants for the 3,124 compounds we extracted. Our domain corpus accounts for variation of most compounds, while only 461 (approximately 15%) of the compounds have no variants. We observe that the most frequent case for more than half of the compounds is that they have two variants. The average number of extracted variants per compound is approximately 1.82.

Variants per compound	Number of compounds
0	461
1	754
2	1,604
3	58
4	69
5	49
≥ 6	128

Table 2: Annotated variants

7. Conclusions

We have introduced a promising method for analyzing term variation in texts, which allows for the semantically grounded detection of variant shapes of a given string set, and with noise tolerated in favor of high recall.

Results have been later refined by applying both ontological restrictions to FE combinations and syntactic restrictions to multiword terms. Tests performed on other technical fields also demonstrate that the method is generalizable at least to domains that show similar conceptualization and standardization traits.

In future work, the integration of a new compound splitting approach into the current method could be tested, with the goal of restricting annotation to those strings which do not violate the splits.

Validation and evaluation steps have been performed in the context of the main study by applying the method to a new corpus and comparing our results with those obtained by other term extraction tools.

8. Bibliographical References

Anić, A. O. and S. K. Zuvela (2017). The conceptualization of music in semantic frames based on word sketches. In 9th International Corpus Linguistics Conference.

Burchardt, A., K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal (2009). Using FrameNet for the semantic analysis of German: Annotation, representation, and automation. In Boas H. C. (Ed.), Multilingual FrameNets in Computational Lexicography, pp. 209-244. De Gruyter Mouton.

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

Corcoglioniti, F., M. Rospocher, and A. P. Aprosio (2016). Frame-based ontology population with pikes. IEEE Transactions on Knowledge and Data Engineering 28(12), 3261–3275.

Evert, S. and A. Hardie (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. http://cwb.sourceforge.net/index.php

Faber, P. (2012). A cognitive linguistics view of terminology and specialized language, Volume 20. Walter de Gruyter.

Faber, P. (2015). Frames as a framework for terminology. Handbook of terminology 1, 14–33.

Fillmore, C. J. and C. Baker (2010). A frames approach to semantic analysis. In B. Heine and H. Narrog (Eds.), The Oxford handbook of linguistic analysis, pp. 313–339. Oxford University Press.

Freixa, J. (2006). Causes of denominative variation in terminology: A typology proposal. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 12(1), 51–77.

Giacomini, L. (2020). Ontology – Frame – Terminology. A method for extracting and modelling variants of technical terms (Forthcoming).

Giacomini, L. (2019). Phraseology in technical texts: a frame-based approach to multiword term analysis and extraction. In Proceedings of Europhras 2019, Santiago de Compostela (ES).

Giacomini, L. (2017). An ontology-terminology model for designing technical e-dictionaries: formalisation and presentation of variational data. In Proceedings of eLex 2017, September 2017, Leiden (NL).

Gómez-Moreno, J. M. U. and M. B. Castro (2017). Semantic and conceptual aspects of volcano verb collocates within the natural disaster domain: A frame-based terminology approach. Cognitive Approaches to Specialist Languages, 330.

Reforgiato Recupero, D., V. Presutti, S. Consoli, A.

Gangemi and A. G. Nuzzolese (2015). Sentilo: Frame-Based Sentiment Analysis. Cognitive Computation 7, pp. 211–225.

Schäfer, J., I. Rösiger, U. Heid, and M. Dorna (2015). Evaluating noise reduction strategies for terminology extraction. In TIA, pp. 123–131.

Schmid, H. and F. Laws (2008). Estimation of conditional probabilities with decision trees and anapplication to fine-grained pos tagging. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, pp. 777–784. Association for Computational Linguistics.

Wüster, E. (1974). Die allgemeine terminologielehre–ein grenzgebiet zwischen sprachwissenschaft, logik, ontologie, informatik und den sachwissenschaften. Linguistics 12(119), 61–106.