Three Approaches to Finding German Valence Compounds

Yannick Versley, Anne Brock, Verena Henrich, Erhard Hinrichs
SFB 833 / Seminar für Sprachwissenschaft
Universität Tübingen
firstname.lastname@uni-tuebingen.de

Abstract

Valence compounds (German: *Rektionskomposita*) such as *Autofahrer* 'car driver' are a special subclass in the otherwise very heterogeneous class of nominal compounds. As the corresponding verb (*fahren* 'to drive' in the example) governs the (accusative) object (*Auto* 'car'), valence compounds allow for a straightforward (event-)semantic interpretation. Hence the automatic detection of valence compounds constitutes an essential step towards a more comprehensive approach to the analysis of compound-internal semantic relations. Using a hand-annotated dataset of 200 examples, we develop an accurate approach that finds valence compounds in large-scale corpora.

1 Introduction

German, Dutch, and other languages exhibit the phenomenon of word formation by compounding: In a process where nouns, verbs and other roots combine with a head noun, a novel word can be formed which is typically interpretable by considering its parts and the means of combination.

Previous research on compounds in German computational linguistics has concentrated on the question of accurately splitting them: Schiller (2005) and Marek (2006) present finite state approaches for accurate compound splitting, Koehn and Knight (2003) use a parallel corpus to find appropriate splits for compounds without oversplitting them.

For the English language, previous years have seen renewed interest in the semantic interpretation of noun-noun compounds which are the most conspicuous kind in English. Research such as that by Girju et al. (2005) and Ó Séaghdha (2007), *inter alia*, has concentrated on automatic classification of the compound-internal relations.

Compounds are a rich source of examples even for semantic relations crossing part-of-speech categories, e.g. when the head part of the compound is a nominalization. In this paper, we want to focus on the detection of one particular kind of cross-part-of-speech relation between nouns and deverbal nominalizations in so-called *valence compounds*. Our goal is to identify noun-object valence compounds among the words occurring in a corpus using a combination of morphological, statistical and semantic evidence.

Valence compounds are an interesting subset of all compounds due to the fact that they have a straightforward (event-)semantic interpretation, namely that the modifier noun fills an argument slot in the event expressed by the head's nominalized verb, as in *Volkszählung* 'people count', which corresponds to an event where the people are counted.

Existing linguistic research cautions us that not everything that looks like the targeted construction (valence compounds with an accusative object modifier) is a valid example: The corpus-based studies of Wellmann (1975) and Scherer (2005) find that, even among *-er* derivations, the resulting noun does not realize an agent in all cases; Kohvakka and Lenk (2007) confirm that agentive nominals can also inherit other kinds of arguments such as prepositional objects (about 10% in their study). Finally, Gaeta and Zeldes

Proceedings of KONVENS 2012 (Main track: poster presentations), Vienna, September 19, 2012

(2012), in their corpus study of valence compounds, remind us that not all such compounds correspond to an interpretable verb-object pair.

Therefore, the kind of valence compounds we are interested in (those that do correspond to an interpretable verb-object pair) constitute only a part of all compounds with a plausible morphology; but it is also the case that, since compounding presupposes a certain degree of semantic integration, such valence compounds are more specific than verb-object collocates in general.

2 Related Work

Lapata (2002) presents an approach to discriminate between compounds with an object- or subject-related modifier and a deverbal nominalization head. From the British Nominal Corpus (BNC), Lapata extracted a 1,277 item sample of such compounds that contained an actual nominalization head according to CELEX or NOMLEX. She then created a gold standard dataset containing a subset of 796 nominalization compounds where the premodifying noun corresponded either to a subject or to an object relation.

To predict whether a given compound is a subject or object nominalization, Lapata estimates the ratio of subject versus object occurrences among the co-occurrences of the noun and verb from which the valence compound is derived, and complements the raw frequency with class-based smoothing via hand-crafted resources (WordNet and Roget's thesaurus), as well as distance-weighted averaging by distributional similarity across verbs. Lapata combines different smoothing methods using decision list classification (Ripper: Cohen, 1996), yielding a final accuracy of 86.1%.

Lapata's work for English has a slightly narrower goal than ours, as she aims to discriminate verb-object nominalization compounds from verb-subject ones, rather than from all other types of compounds as is our goal. Nonetheless, her work is most similar in spirit to the specialized approach to valence compounds that we will advocate later in this paper.

3 Dataset

The target set of compounds for our experiments consists of 100 compound instances labeled as valence compounds and 100 others. To find both the positive and negative examples, we prepared a list of nouns exhibiting a morphological structure compatible with being a valence compound.

Using the morphological analyzer SMOR (Schmid et al., 2004), we prepared a list of such compounds in which the nominalized verbs and nominal non-heads are simplex words (i.e., not compounds or derivations). Specifically, the SMOR analysis had to consist of a bare noun followed by a bare verb (in contrast to *Dienstagabend* 'Tuesday night', which does not have a deverbal head, or to *Netznutzungsentgelt* 'network access fee', which has a complex modifier *Netznutzung* 'network access'). The dataset includes a variety of nominalization suffixes: *-ung, -en, -er, -erei,* as well as some less frequent ones.

A compound is labeled as a valence compound when its most common interpretation is that of a verb and its direct object, although some of them may be ambiguous. As a negative example, the word *Serienmörder* 'serial killer' consists of the nominalized verb *morden* 'to murder' and the noun *Serie* 'series'. While the word could be used to denote a TV executive killing a series (i.e., canceling a show), the common meaning is different and we would not count the word as a valence compound.

4 Methods

As a text collection that provides contexts for the words or word pairs that interest us, we use the *TüPP-D/Z* corpus of `tageszeitung` news articles from 1986 to 1999 (Müller and Ule, 2002), and the *web-news* corpus, a 1.7 billion word collection of online news articles by Versley and Panchenko (2012). The parsing model used in the pipeline is based on MALTParser, a transition-based parser (Hall et al., 2006), and uses part-of-speech and morphological information from RF-Tagger (Schmid and Laws, 2008) as input. Using the MALTParser support for linear classification by Cassel (2009), we reach a parsing speed of 55 sentences per second.

4.1 Simple Association Statistics

One very straightforward idea for the identification of valence compounds is to check for a valence of the verb (corresponding to the deverbal

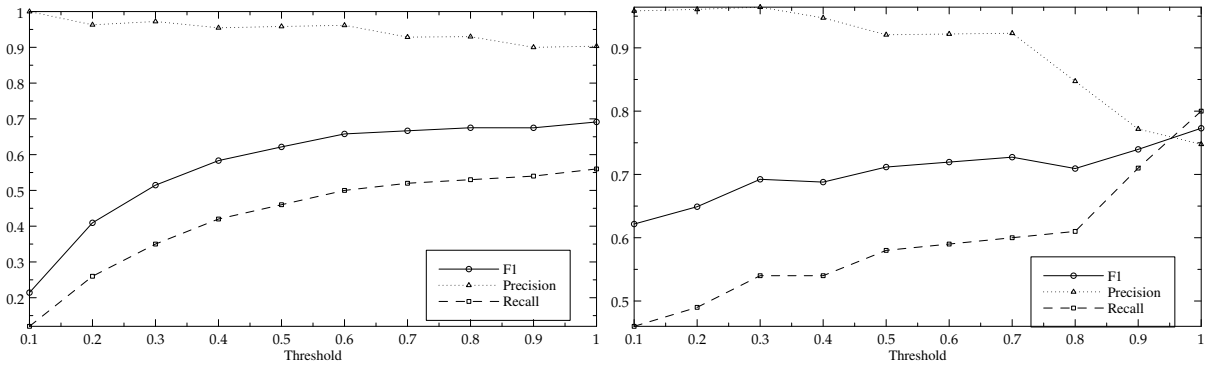


Figure 1: Precision/Recall/F_1 values for different thresholds on pointwise mutual information (left) and log-likelihood (right)

SVM	Acc	F_1
GN+triples+w1w2	0.695	0.697
GN+triples/scale	0.705	0.619
GN	0.615	0.617

Table 1: Results for feature-rich SVM classification

part of a putative valence compound) with a selectional preference that would admit the noun (i.e., the premodifying part of a candidate).

From the verb-object pairs in our corpus (i.e., accusative objects tagged as OBJA), we calculate association statistics that consider counts for this relation across different verbs and arguments seen in the corpus. The most common statistics are *pointwise mutual information* and conservative estimates thereof, and the G^2 significance statistic (*log-likelihood*) proposed by Dunning (1993).

Figure 1 shows precision and recall for the task of identifying valence compounds when using different thresholds on the respective statistic.

Another way to look at the problem is to look at the *relative frequency* of a verb and a noun that co-occur in a sentence being connected by an OBJA edge (or the subject of a passive construction) rather than by some other dependency (e.g., the noun being the subject of the verb or occurring inside a prepositional phrase). As seen in Figure 2, imposing different thresholds on the ratio between object and non-object occurrences yields a recall between none and about 80% of valence compounds, and precision above 75%.

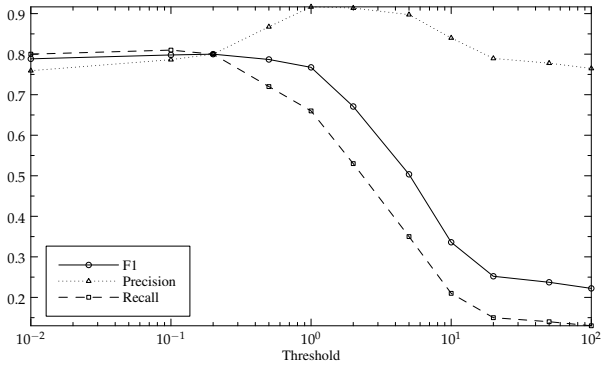


Figure 2: Precision/Recall/F_1 values for different thresholds on relative frequency

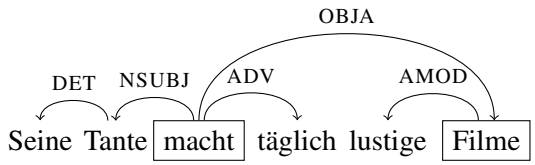
4.2 Feature-rich Supervised Classification

In the domain of more general approaches to the prediction of a relation between two words – for example, the two nouns in a noun-noun compound – more feature-rich approaches have been developed that can take into account *all paths* that occur between the two target words as well as *taxonomic* information.

In our version of such an approach, we used taxonomic relations in GermaNet (Kunze and Lemnitzer, 2002; Henrich and Hinrichs, 2010), denoted as *GN* in Table 1, as well as features derived from corpus co-occurrences, namely *triples* along the dependency path as well as words occurring in-between or around the target words in a co-occurrence (*w1w2*). The corpus-based features are illustrated in Figure 3.

4.3 Decision-tree Based Combination

To combine multiple statistics such as those described in subsection 4.1 and possibly smoothed



His aunt makes daily funny movies
"his aunt makes funny movies every day"

`w1w2`	Seine$_{w_2,w_1}$ Tante$_{w_2,w_1}$ $_{w_2}$täglich$_{w_1}$
`triples`	$w_1 \uparrow_{\text{OBJA}} w_2$

Figure 3: Features for SVM classification

JRip	Acc	F_1
mi+ll	0.770	0.776
rel	**0.860**	**0.863**
mi+ll+rel	0.840	0.837
mi+ll/avg	0.740	0.743
rel/avg	0.845	0.841
mi+ll+rel/avg	0.840	0.833
J48	Acc	F_1
mi+ll	0.785	0.792
rel	**0.860**	**0.863**
mi+ll+rel	0.850	0.851
mi+ll/avg	0.765	0.759
rel/avg	0.850	0.854
mi+ll+rel/avg	0.845	0.846
AdaBoost+J48	Acc	F_1
mi+ll	0.780	0.786
rel	0.830	0.835
mi+ll+rel	0.810	0.812
mi+ll/avg	0.780	0.770
rel/avg	0.710	0.655
mi+ll+rel/avg	0.830	0.830

Table 2: Results for decision tree classification

variants thereof, we use symbolic learning techniques such as decision lists (Cohen, 1996) or decision trees (Quinlan, 1993) which are aimed at finding logical combinations of thresholds, possibly in combination with AdaBoost as a meta-learner (Freund and Schapire, 1997).

In our case, we try both the association statistics for the `OBJA` relation and the relative frequency heuristic, as either raw values or averaged over a group of 10 related nouns. The set of related nouns was determined by taking 30 surrounding terms from GermaNet, then ranking by distributional similarity (using premodifying adjective and accusative object collocates, and similarity based on Jensen-Shannon divergence).

5 Results and Discussion

The statistics in Figures 1 and 2 have been determined on the full dataset, as the risk of overfitting is very low when fixing a single threshold parameter. The more comprehensive approaches using high-dimensional features (Subsection 4.2) or combining multiple statistics using symbolic learning (Subsection 4.3) were run as ten-fold crossvalidation where each 10% slice was predicted using a classifier built from the remaining 90% of the data.

The statistics show that the relative frequency heuristic yields the best F_1 measure (0.80) for a 1:5 cutoff ratio between accusative objects and other paths. In contrast, association statistics yield a best F_1 of 0.77 (log-likelihood).

In comparison, the feature-rich approach (see results in Table 1) does not seem particularly attractive: taxonomic information from GermaNet alone yields an F_1 measure of 0.62, and even the most sophisticated feature set that additionally takes into account surface-based and dependency-based features from the collocation contexts of noun and verb only yield an F_1 measure of 0.70.

The decision list and decision tree based methods allow to explore (and potentially to combine) larger sets of features with different corpora and smoothed/unsmoothed variants of the statistics. In our experiments, we found that the best classifier selected the relative frequency heuristic (but selecting statistics from the larger *web-news* corpus instead of the smaller TüPP-D/Z ones), reaching an F_1 measure of 0.86.

6 Summary

We demonstrated several methods to analyze the relations inside nominalizations and identify valence compounds. While a generic feature-rich approach works moderately well, we find that tightly focused statistics such as those investigated by Brock et al. (2012) can be easily combined using symbolic machine learning methods to yield a highly accurate discrimination between valence compounds and non-valence compounds.

Acknowledgements We are grateful to the three anonymous reviewers for insightful comments. Anne Brock and Yannick Versley were supported by the DFG as part of SFB 833.

References

Brock, A., Henrich, V., Hinrichs, E., and Versley, Y. (2012). Automatic mining of valence compounds for German: A corpus-based approach. In *Digital Humanities Conference Abstracts*, Hamburg. Hamburg University Press.

Cassel, S. (2009). MaltParser and LIBLINEAR - transition-based dependency parsing with linear classification for feature model optimization. Master's thesis, Uppsala University.

Cohen, W. W. (1996). Learning trees and rules with set-valued features. In *Proc. 13th National Conf. on Artificial Intelligence (AAAI 1996)*.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Freund, Y. and Schapire, R. (1997). Decision-theoretic generalization of on-line learning and an application to boosting. *J Comp Sys Sci*, 55(1):119–113.

Gaeta, L. and Zeldes, A. (2012). Deutsche Komposita zwischen Syntax und Morphologie: Ein korpusbasierter Ansatz. In Gaeta, L. and Schlücker, B., editors, *Das Deutsche als kompositionsfreudige Sprache: Strukturelle Eigenschaften und systembezogene Aspekte*, pages 197–217. De Gruyter, Berlin.

Girju, R., Moldovan, D., Tatu, M., and Antohe, D. (2005). On the semantics of noun compounds. *Journal of Computer Speech and Language - Special Issue on Multiword Expressions*, 19(4):479–496.

Hall, J., Nivre, J., and Nilsson, J. (2006). Discriminative classifiers for deterministic dependency parsing. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 316–323.

Henrich, V. and Hinrichs, E. (2010). GernEdiT - the GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235.

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *EACL 2003*.

Kohvakka, H. and Lenk, H. E. H. (2007). Streiter für Gerechtigkeit und Teilnehmer am Meinungsstreit? Zur Valenz von Nomina Agentis im Deutschen und Finnischen. In *Wahlverwandtschaften. Valenzen - Verben - Varietäten*, pages 195–218. Georg Olms, Hildesheim, Zürich, New York.

Kunze, C. and Lemnitzer, L. (2002). GermaNet – representation, visualization, application. In *Proceedings of LREC 2002*.

Lapata, M. (2002). The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.

Marek, T. (2006). Analysis of German compounds using weighted finite state transducers. Bachelor of arts thesis, Universität Tübingen.

Müller, F. H. and Ule, T. (2002). Annotating topological fields and chunks – and revising POS tags at the same time. In *Proc. 19th Int. Conf. on Computational Linguistics (COLING 2002)*.

Ó Séaghdha, D. (2007). Annotating and learning compound noun semantics. In *Proceedings of the ACL07 Student Research Workshop*.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Scherer, C. (2005). *Wortbildungswandel und Produktivität: Eine empirische Studie zur nominalen* -er-*Derivation im Deutschen*. Niemeyer.

Schiller, A. (2005). German compound analysis with *wfsc*. In *Proceedings of the 5th International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP 2005)*.

Schmid, H., Fitschen, A., and Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of LREC*.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *COLING 2008*.

Versley, Y. and Panchenko, Y. (2012). Not just bigger: Towards better-quality Web corpora. In *Proceedings of the 7th Web as Corpus Workshop (WAC-7)*, pages 44–52.

Wellmann, H. (1975). *Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache.* Schwann, Düsseldorf.