Evaluating the Morphological Compositionality of Polarity

Josef Ruppenhofer and Petra Steiner
Institute for German Language
R5, 6-13
D-68161 Mannheim, Germany
ruppenhofer|steiner@ids-mannheim.de

Michael Wiegand
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
michael.wiegand@lsv.uni-saarland.de

Abstract

Unknown words are a challenge for any NLP task, including sentiment analysis. Here, we evaluate the extent to which sentiment polarity of complex words can be predicted based on their morphological make-up. We do this on German as it has very productive processes of derivation and compounding and many German hapax words, which are likely to bear sentiment, are morphologically complex. We present results of supervised classification experiments on new datasets with morphological parses and polarity annotations.

1 Introduction

The vast variety of language that speakers use to express evaluations presents a key challenge for sentiment analysis. Any approach to sentiment analysis has to grapple with problems of coverage. Coverage gaps can arise for a number of different reasons, ranging from typos, foreign language material used in code-switching contexts, to rare words. We are interested in the latter sources for coverage gaps. They present a particular problem because, by definition, rare words cannot readily be modeled using a data-driven, corpus-based approach because of the lack of training instances. However, sub-word analysis may give us a significant hook into analyzing such rare words. In this paper, we evaluate the hypothesis that we can model the polarity of German words based on the properties of their morphological make-up, which has parallels to recent work on sentiment composition on the syntactic level (Socher et al., 2013; Haas and Versley, 2015). German has a relatively rich morphology and very productive processes of derivation and compounding. For instance, Baroni et al. (2002) report that 83% of the compounds in a 28-million word German newswire corpus had a corpus frequency of 5 or less. Since many compounds are hapaxes and since prior work has shown that hapax words are often subjective (Wiebe et al., 2004), it would be particularly desirable to be able to infer the polarity of compounds from their components.

To that end we conduct extensive experiments on a data set with complete morphological parse trees that we compiled from existing resources and which we augmented further.[1] This data set is presented in §2. We lay out the features and the setup of our main experiments in §3. Their results are discussed in §4 where we show the efficacy of psycholinguistic ratings as features and also demonstrate that a shallower analysis at the level of immediate constituents is better than 'going deep' to the leaf level. We present an additional set of experiments in §5 that shed light on the limits of our results from the main experiment when we attempt to generalize to very different data. We discuss related work in section §6 before concluding in §7.

2 Data

Our data mainly consists of lemmas sampled out of the PolArt polarity dictionary (Klenner et al., 2009), a manually curated resource with around 8000 entries, found to be of high precision by (Emerson and Declerck, 2014). Most entries are subjective words but PolArt also includes a few intensifiers (INT)[2], and shifters (SHI). Each subjective entry is labeled as positive (POS), negative (NEG) or neutral (NEU). The polar entries have a discrete score reflecting the degree of their positivity or negativity.

As PolArt has only 55 neutral entries and is

[1] Annotations and morphological analyses are available at https://github.com/josefkr/morphcomp

[2] PolArt's intensifiers also include downtoners such as *kaum* 'barely'

Proceedings of Recent Advances in Natural Language Processing, pages 625–633,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_081

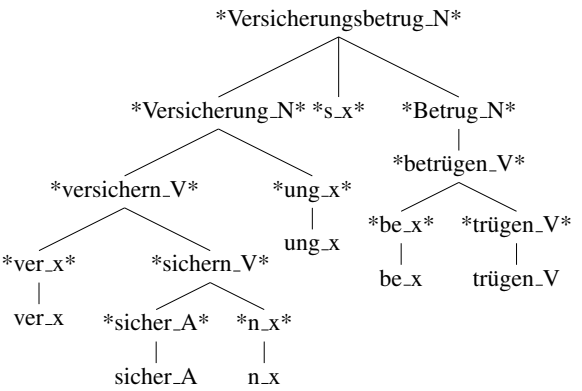

Figure 1: Morphological sample parse for *Versicherungsbetrug* 'insurance fraud'

short on compounds, we added words found in the CELEX database for German (Baayen et al., 1993). CELEX contains information about orthography, phonology, morphology, word class, argument structure as well as word form and lemma frequencies for words in German, Dutch and English. The use of CELEX is motivated by the fact that this resource contains among its morphological information not only segmentations but also complete human-created hierarchical parses, which we use in our experiments. We added polarity annotations for the CELEX entries.

Figure 1 shows the morphological parse for a German compound noun. It contains both components that are themselves words and as such (potentially) have entries in PolArt (e.g. the verb *versichern* 'insure') as well as components that are bound morphemes such as the nominalizing derivational suffix *-ung* or the linking element *s*. We expanded PolArt with entries for 167 affixes, assigning them polarity and intensity values.

Of the total 9300 lemmas, we use 8400 for training and testing in a 10-fold cross-validation setting. While we do not tune parameters for the classifiers, we set aside a development set of 900 lemmas for feature tuning and error analysis.

2.1 Statistics

We use 5937 PolArt entries and 2443 items added from CELEX for training and testing. The CELEX-derived entries are mostly neutral. As shown by Table 1, the distribution of POS (N=noun, A=adjective, V=verb, R=adverb, P=preposition) is dominated by the major lexical classes: nouns, verbs and adjectives.

2.2 Agreement

To test if our polarity annotations are compatible with those of PolArt's creators, we re-labeled a random sample of entries. We also labeled

		Parts of speech					
		N	A	V	R	P	Total
POS	PolArt	886	994	419	0	0	2299
	CELEX	197	57	7	2	0	263
NEG	PolArt	1572	1267	737	0	0	3576
	CELEX	354	78	20	5	1	458
NEU	PolArt	0	53	2	0	0	55
	CELEX	1629	86	6	0	0	1721
SHI	PolArt	0	4	3	0	0	7
	CELEX	0	0	0	0	0	0
INT	PolArt	4	14	2	0	0	20
	CELEX	0	1	0	0	0	1
Total		4642	2554	1196	7	1	8400

Table 1: Distribution of POS and polarity

words from our CELEX expansion set to ensure that annotation was consistent on the new data (cf.§2.2.1). As we use novel information on derivational affixes in our experiments, we did a second annotation study (cf.§2.2.2). We sampled 100 derivational morphs. In cases of allomorphy (e.g. the negative prefix *in/im/ir/il-*), we kept only one variant. This cut down our set to 85 instances.

2.2.1 Agreement on Words

A student assistant (A1) and one of the authors (A2) labeled 200 roots for polarity, 100 from PolArt and 100 from CELEX. The kappa scores for all words and the two subsets are given in Table 2. Agreement is high on PolArt as it basically contains only polar words (cf. Table 1). On the added CELEX data, it is high, too, since that set contains many clearly neutral cases such as nouns referring to concrete objects. Agreement is reasonably high, both between the annotators and in comparison to PolArt. Since A2 has higher agreement with PolArt, we use A2's labels when A1 and A2 disagree.

2.2.2 Agreement on Affixes

Observed percent agreement on affixes is 0.85, the Cohen's kappa value is 0.67. The class distribution among affixes is significantly different to roots and words, with neutral affixes accounting for 67% of the affixes types. The disagreements for affixes centered on two distinctions unrelated to the neutral class: SHI vs POS/NEG and INT vs POS/NEG. To illustrate the former, one annotator treated *-frei* 'free of/without' as a shifter, whereas the other treated it as an affix with positive polarity. An example of the latter is *-reich* 'abounding with' which one annotator treated as an intensifier and the other as having positive polarity.

Pair	PolArt	CELEX	All
A1 vs. A2	0.95	0.84	0.90
A1 vs. PolArt	0.72	n/a	n/a
A2 vs. PolArt	0.80	n/a	n/a

Table 2: Kappa measures on all data and subsets

3 Experiments

In our classification experiments we assign a positive, neutral or negative polarity to a composite word based on its morphological make-up. With shifters and intensifiers being so rare (cf. Table 1), we leave them out of consideration.

3.1 Feature Groups

We first discuss the features that we use. They may apply to the composite word, to its immediate constituents (ICs), or to the leaves. In Figure 1, the ICs are *Versicherung*, *s* and *Betrug*.

3.1.1 Structural Features

This group of features considers information about the structure of the word as a whole.

height the height of the morphological parse tree

nt count the number of non-terminal nodes[3]

leaf count the number of leaves on the parse tree

compound is the word a compound formed by its immediate constituents? E.g. *Angsthase* 'scaredy cat' (lit. 'fear bunny/rabbit') is a compound at the level of immediate constituents. The adjective *angsthasenhaft* 'scaredy cat-like' contains this compound but is itself a derived form at the IC level.

category change does the composite word belong to a different POS category than all of its immediate constituents?

3.1.2 POS Features

We use three sets of features representing POS information on three levels. Each set consists of five indicator variables, one for each POS.

comp pos part of speech of the composite word

ic pos POS present among the ICs

leaf pos POS present among the leaves

3.1.3 Psycholinguistic Features

If possible, we extract psycholinguistic ratings for the whole word, its immediate constituents and the leaves. These features have been successfully used in the task of identifying metaphor, which also has a significant expressive function (Turney et al., 2011; Gargett et al., 2014). Psycholinguistic ratings were also explored for the task of polar intensity scoring by Ruppenhofer et al. (2014).

The first dimension places words on a scale from abstract to concrete (**abstconc**). Abstract words denote things we do not perceive directly (*integer*, *politics*, ...) whereas concrete words refer to things we can perceive (*sound*, *scent*, ...). The second feature concerns imageability. A large subset of concrete words have a high imageability (**img**). These words refer to things that we can actually see (*chestnut*, *police jeep*, ...). The third rating dimension is valence (**val**), which measures the pleasantness of a word (*gift* vs. *punishment*). The **arousal** dimension represents the intensity of emotion caused by a stimulus (*alert* vs *calm*).[4]

Our affective ratings derive from the work of Köper and Schulte im Walde (2016). While the creation of this resource involved some automatic translation from two English resources (Brysbaert et al. (2014); MRC database[5]) as well as score harmonization, it holds information for 350k words and is thus far more comprehensive than the affective norm data of Kanske and Kotz (2010) or Lahl et al. (2009). It is also much larger than common polarity lexicons such as PolArt (Klenner et al., 2009) or GermanPolarityClues (Waltinger, 2010).

3.1.4 Polarity Features

We use polarity information for immediate constituents and leaves. We take it from PolArt itself or our own supplemental annotations. We define one set of count features (**ic pol**) for the number of immediate constituents that are (i) positive, (ii) negative, or (iii) neutral, (iv) intensifying or downtoning, or (v) negating. We define a parallel group of features (**leaf pol**) for the leaves. Both feature sets include two more features representing the minimum and maximum polarity values present.

3.1.5 Lexical Features

One group of lexical features captures the presence or absence of individual derivational affixes at the level of immediate constituents (**ic affix**). Another set represents the presence of frequent lexical words or morphemes among the ICs (**ic lex**). An item must occur in at least 5 composite words to qualify as a lexical feature.

3.2 Classifiers

We employ a set of standard classifiers, Naive Bayes (NB), logistic regression, and linear SVM in the implementations of scikit-learn (Pedregosa

[3] This feature is highly correlated with height.

[4] Valence and arousal are part of (Osgood et al., 1957)'s well-known theory of emotions.

[5] http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm

```
1: pol="NEU"
2: if pos_ct > neg_ct then
3:     pol="POS"
4: else if pos_ct < neg_ct then
5:     pol="NEG"
6: else if pos_ct == 0 and neg_ct == 0 then
7:     pass
8: else if pos_ct > 0 and neg_ct > 0 then
9:     pol="NEG"
10: else
11:     pass
12: end if
13: if shi_ct%2 == 0 then
14:     pass
15: else
16:     if pol == "NEG" then
17:         pol="POS"
18:     else if pol == "POS" then
19:         pol="POS"
20:     else
21:         pol="NEG"
22:     end if
23: end if
```

Figure 2: Baseline algorithm

et al., 2011). Our evaluation setup is 10-fold cross-validation. We compare different classifiers using McNemar's test with continuity correction (Japkowicz and Shah, 2011). Its χ^2 statistic depends on the number of cases where the classifiers disagree but one matches gold. We apply the test to the concatenated predictions across all folds.

3.2.1 Rule-based Polarity Baseline

A weak majority-class baseline would have a low accuracy of 48.2% on train/test. We define a much stronger rule-based baseline using simple heuristics about polarity and shifting. The algorithm is portrayed in Figure (2). It consists of two steps. The first condition block (lines 1-11) produces an initial polarity for the whole word based on the quantitative relations between positive, negative and neutral morphemes. The second block (12-20) potentially modifies the initial polarity, if there is an odd number of shifters present (12). A negative bias is built in in two places. When positive and negative items occur in equal non-zero number, we default to negative (8-9). When an odd number of shifters occurs only with neutral items, we still assign negative polarity (19-20).

4 Results

We first report basic three-way classification results for training and testing on all multimorphemic words with our best setting. The results in Table 3 show that the SVM and logistic regression classifiers can beat the baseline in terms of accuracy, while Naive Bayes cannot. Considering the F1-score for the individual classes (+ = pos, ∼ = neu, − = neg), we see that while there is better performance for the largest class (negative) the classifiers do learn the properties of the less frequent classes quite well, too. The linear SVN and logistic regression classifiers differ little but Naive Bayes is consistently lower-performing. As classifier optimization is not our focus, we henceforth only use the logistic regression classifier.

classifier	data	feats.	Acc.	+ F1	∼ F1	- F1
baseline	all	all	0.74	0.70	0.67	0.81
Naive Bayes	all	all	0.74	0.70	0.76	0.77
SVM	all	all	0.84$^{\diamond}$	0.80	0.84	0.87
logistic	all	all	0.85$^{\bigstar}$	0.81	0.86	0.88

statistical significance testing: McNemar's; ★ better than SVM at p< 0.001; ◇ better than Naive Bayes at p< 0.001

Table 3: 3-way classification performance on multi-morphemic words on train/test (10-fold CV)

The results on the dev set (not shown for lack of space) pattern like those for the train/test-data.

4.1 Results per POS

Table 4 shows accuracy per POS on the train/test set, comparing the logistic regression classifier to our negatively-biased baseline. For nouns the improvement over the baseline is smallest, which is no surprise as the baseline is strongest for nouns. The biggest gains are seen for verbs.

	N	V	A	all
baseline	0.79	0.59	0.69	0.74
logistic	0.87	0.83	0.84	0.85

Table 4: 3-way classification accuracy per POS on train/test

4.2 Compounding vs Derivation

Table 5 shows, somewhat surprisingly, that compounds seem to be no harder to handle than derived forms for the logistic regression classifier, which is significantly better than the baseline for both sets. For the baseline, there is more of a gap.

data	classifier	**Acc.**	**+ F1**	**∼ F1**	**- F1**
compounds	baseline	0.80	0.74	0.85	0.78
compounds	logistic	0.86$^{\flat}$	0.72	0.90	0.85
derived forms	baseline	0.72	0.71	0.53	0.82
derived forms	logistic	0.85$^{\flat}$	0.82	0.79	0.88

statistical significance testing: McNemar's; ♭ better than baseline at p< 0.001;

Table 5: Average 3-way classification performance on train/test

4.3 Impact of Feature Groups

Table 6 shows the effect of using only particular feature groups when carrying out 3-way clas-

sification in a cross-validation setting on the dev set. The dev set contains 460 negative words, 153 neutral items and 284 positive ones.[6] As expected, overall performance is a bit lower than on the much larger train/test set.

Polarity information about leaves (leaf pol) is useful but information on immediate constituents (ic pol) is more so. Combining the two does not improve over IC-level features alone. Affective ratings (psych) are also helpful. Affective ratings for the composite word (comp psych) are very valuable. The valence dimension by itself has the greatest impact while the other dimensions (arousal, abstconc and img) are not very predictive individually. However, the full set of ratings is markedly better than valence alone. Affective ratings below the level of the composite word (leaf psych, ic psych) still carry significant information. The pattern observed for polarity repeats here: information at the IC-level is better than information about leaves. Combining the two gives no significant boost. More generally, relying on the combination of all features at the level of ICs (ic all) is better than using the features at the leaf level (leaf all). This concords with the observation that structural features are not very predictive.

features	**Acc.**	**+ F1**	**∼ F1**	**- F1**
all	0.83★	0.79	0.81	0.86
structural	0.50	0.00	0.00	0.66
lexical	0.52	0.38	0.31	0.64
leaf pos	0.56	0.09	0.58	0.68
ic pos	0.51	0.00	0.48	0.65
ic & leaf pos	0.55	0.12	0.57	0.68
comp pos	0.50	0.00	0.00	0.66
ic pol	0.74	0.70	0.61	0.84
leaf pol	0.65	0.58	0.46	0.77
ic & leaf pol	0.74	0.69	0.61	0.83
ic psych	0.64	0.55	0.57	0.72
leaf psych	0.60	0.49	0.52	0.68
ic & leaf psych	0.64	0.55	0.58	0.71
comp psych	0.76	0.73	0.59	0.81
comp val	0.68	0.68	0.00	0.79
comp arousal	0.49	0.00	0.26	0.66
comp imageability	0.50	0.00	0.00	0.66
comp abstconc	0.50	0.00	0.05	0.67
ic all	0.80◇	0.76	0.77	0.85
leaf all	0.73	0.67	0.69	0.78
ic & leaf all	0.80	0.75	0.78	0.83

statistical significance testing: McNemar's; ★ better than *all ic & leaf* at p< 0.001; ◇ better than *all leaf* at p< 0.001

Table 6: Average 3-way classification performance on the dev set for multi-morphemic words, omitting some features

[6]The set also contains 3 intensifiers but we leave these aside as we perform only a three-way classification.

4.4 Error Analysis

Table 7a shows a confusion matrix for three-way classification, summed across all 10 folds. The confusions between the polar classes (neg for pos, or pos for neg) make up most (51%) of the errors. This picture seems, however, to be due to the derived forms but not the compounds. As Table 7b shows, for compounds the 'fatal' confusions between the two polar classes are much less frequent (17%). Instead, we observe many confusions of what is actually polar as neutral (59%).

	pos	neu	neg
pos	**1741**	143	288
neu	117	**1547**	110
neg	242	138	**2936**

(a) Derivations and compounds

	pos	neu	neg
pos	**159**	64	20
neu	21	**889**	37
neg	20	78	**441**

(b) Compounds only

Table 7: Confusion matrices for train/test

Inspection of the predicted and gold labels on the dev set suggests the key source of error is that our lemma-based approach cannot deal with polysemy and the effects of idiosyncratic lexicalization. One example of this is the complex noun *Blinddarm* 'appendix' (lit. 'blind intestine'). The first component of this noun is the adjectival root *blind* 'blind' which is listed in PolArt as negative. The negative polarity of the adjective is however irrelevant to the rare meaning of 'lacking an opening' which *blind* has in *Blinddarm*. Similarly, the derived verb *umsorgen* was tagged as negative due to a PolArt entry that reflects a negative sense of 'worry' for *sorgen*, whereas the prefix *um-* is treated as neutral. Within *umsorgen*, however, *sorgen* occurs in its positive meaning of 'care (for)'. A third illustrative example is the derived adjective *lachhaft* 'laughable', which was predicted to be positive since the root *lach-* 'laugh' has positive polarity while the suffix *-haft* is neutral. But like its English gloss, *lachhaft* is negative.

These issues have clear parallels in the syntactic domain with multi-word expressions, which are well known to pose problems of compositionality (Sag et al., 2002). And in fact, most German compounds translate to English compounds which are considered a core part of the typology of multi-word expressions (Schneider et al., 2016). Likewise many German prefix verbs correspond to English particle verbs, which are MWEs in English.

5 Out of Domain Testing for Compounds

We suspect that the conclusions we can draw based on our PolArt and CELEX data may not carry over to other types of vocabulary. Con-

sider that 8226 (88.4%) out of our total 9300 items are also listed in GermaNet (Hamp and Feldweg, 1997), the German counterpart of WordNet, and 93% of the items are contained in dlexdb, a lexical database for psycholinguistic research (Heister et al., 2011). These words are thus rather frequent and common. We now want to look at less common words and at compounds in particular. While the results for compounds on PolArt suggested that they were potentially as easy to handle as derived forms, that result was counter-intuitive: compound interpretation usually poses considerable challenges. The other reason to focus on compounds is that compounds are often hapaxes and that hapaxes in turn often express subjectivity.

We add data from two sources: from the collaborative online lexicon Wiktionary and from Wortwarte[7]. Wortwarte ('word watch') is a project that aims to extract neologisms from web-data, mostly from online newspapers. Words from these sources are less likely to be covered by GermaNet or dlexdb than PolArt entries (cf. Table 8).

	Wiktionary	Wortwarte
sample size	200	100
pos/neu/neg	3/38/159	9/53/38
% in GermaNet	36	1
% in dlexdb	68	1

Table 8: Wiktionary and Wortwarte data

So that the new items are not at a disadvantage relative to the PolArt and CELEX-derived items, we augment our polarity resource with new annotations for any unseen component words and morphemes introduced by the Wiktionary and Wortwarte data and also produce morphological parses as needed. With both data sets, we use the best system configuration found on the train/test data.

5.1 Wiktionary Compounds

The Wiktionary compounds were extracted by looking for lemmas labeled as compounds, with at least one sense marked as positive or negative/pejorative. We selected a random sample of 200 items. We re-annotated polarity as we consider the items as lemmas and a negative or positive word sense may not be salient to an annotator making a lemma-level judgment. The polarity distribution among the items is very biased (cf. Table 8).

On Wiktionary, the polarity baseline achieves an accuracy of 0.40. Unsurprisingly, it does worst on the few positive instances but for the neutral and negative classes the F1-scores are not far apart (0.38 vs 0.44). The gravest source of error are actual negatives predicted as neutral. This results from the fact that the sample contains many cases such as *Quotenfrau* 'token woman' (lit. quota woman), which lack polar components but have a negative connotation. With an accuracy of 0.49, our system outperforms the baseline. The difference is significant at the α=0.01-level according to a McNemar's test. Like the baseline, our system does worst on the positive class. However, it handles the negative class significantly better than the rule-based baseline. Still, the same overall picture holds: the main source of error are negative instances predicted as neutral (cf. Table 9a).

	pos	neu	neg
pos	**0**	3	0
neu	1	**31**	6
neg	8	85	**66**

(a) Wiktionary

	pos	neu	neg
pos	**7**	1	1
neu	20	**18**	15
neg	6	7	**25**

(b) Wortwarte

Table 9: Confusion matrices for Wortwarte and Wiktionary compounds

On the Wiktionary data there are hardly any confusions between the polar classes. In this regard, the Wiktionary compounds are like PolArt's compounds and unlike PolArt's derived forms.

5.2 Wortwarte Compounds

Many Wortwarte items are of a domain-specific technical nature (e.g. *Ankertaumine* 'tethered mine') while others are playful (*Egoaufbauprogramm* 'ego-boosting program'). Typically, the items remain low-frequency and do not enter the general lexicon (cf. Table 8). As with Wiktionary, the classification results for Wortwarte are low: we achieve an accuracy of 0.50. The confusion matrix (cf. Table 9b) shows that many actually neutral items are predicted as positive or negative. The low accuracy is not due to missing polarity information: recall that we augmented our lexicon with this information for any unseen items. Two factors seem to be at play. On the one hand, Wortwarte includes words that are more creative and less predictable than the words in PolArt, and on the other hand many more neutral items than are found in the polarity lexicon. Note that the biased rule-based baseline outperforms our system significantly (according to McNemar's test), achieving an accuracy of 0.73, mainly due to much better recognition of neutral cases. Comparing the confusion matrix for Wortwarte to those for Wiktionary and PolArt (Table 7a) suggests that the

[7] http://www.wortwarte.de/

Wortwarte data represents its own kind of challenge.

6 Related Work

Moilanen and Pulman (2008) carried out a study on English that explored how well it was possible to classify unknown English words into one of three polarity classes based on morphological analysis that mainly considered affixation and zero conversion. Our work is different in several respects. First, we focus exclusively on morphology, whereas Moilanen and Pulman (2008) also used information about the syllable structure of words. Second, Moilanen and Pulman (2008) evaluated on a combination of infrequent words from the BNC and what they called 'junk' entries from a web-corpus. We did not use entries of the latter sort as we did not want to mix the issue of normalization into our setup. While we did test on low-frequency words, we purposefully also used high-frequency words to investigate the differences in compositionality in between words of different frequency bands. Because we used a polarity lexicon, we only have citation forms, whereas Moilanen and Pulman (2008) used inflected word forms. Finally, because the division of labor between morphology and syntax differs between German and English, our data included many cases that in English would be encountered as multi-word expressions.

Neviarouskaya et al. (2011) use morphological knowledge about derivation and compounding to perform a rule-based expansion of a base polarity lexicon. Newly proposed formations are added only if they are listed in WordNet (Miller, 1995). By comparison, our work is learning-based, unrestricted in terms of the morphological rules by which composite words are built up, and it addresses neutral items in addition to polar ones. Our results are consistent with the work of Neviarouskaya et al. (2011): entries in a human-curated sentiment lexicon are likely to be sufficiently compositional so that an expansion based on derivation and restricted compounding (within the bounds of a general language resource such as WordNet) makes sense.

In other related work, Wiegand et al. (2016) developed an approach to classify the modifiers of German compounds as expressing either the source or target of evaluation, or neither. For compounds whose head words are not always subjective, they use distributional similarity to first classify the compounds as a whole as being either subjective or not. Their setting assumes that the compounds in question are reasonably frequent to make distributional information reliable. In our research, we are interested in the predictive value of word-internal information by itself, especially with an eye towards handling rare words.

7 Conclusion

We presented a learning-based approach to the task of predicting the polarity of German words from their morphological make-up, focusing on derived forms and compounds in a new data set that we compiled and which we will make available. Using knowledge about the polarity of components and information about polarity shifting morphemes, we achieved a maximum performance of 85% accuracy on the train/test-set derived from the PolArt and CELEX lexical resources. Our results showed that psycholinguistic ratings for affective dimensions, even automatically generated ones as those of Köper and Schulte im Walde (2016), can substitute for or be combined with polarity features. The experiments also demonstrate that information at the immediate constituency level is more reliable than those at the level of leaves. As other structural features did not have a large impact, it seems advisable to proceed top down and only as far as is necessary to find known components. The experiments further suggested that derivations and compounds both seemed amenable to morphological analysis for the purposes of polarity prediction.

However, experiments on data from Wiktionary and Wortwarte showed that the knowledge we learned on the augmented PolArt resource does not readily transfer to compounds that are rarer and either more domain-specific, colloquial and/or more playful. Lexicalized compounds such as *Zeitungsente* 'canard' (lit. 'newspaper duck') which contain no polar part do not seem to be well represented in PolArt and they most likely cannot be handled sufficiently well on the basis of morphological knowledge alone. In future work, we therefore want to focus on a) using local context to analyze particular instances of compounds and on b) using corpus-derived information about the polarity of a given sub-word unit across multiple complex words it occurs in. The latter seems a promising addition to using polarity information about the uses where the items occur as free words.

Acknowledgments

The authors were partially supported by the German Research Foundation (DFG) under grants RU 1873/2-1 and WI 4204/2-1.

References

R. Harald Baayen, R. Piepenbrock, and H. van Rijn. 1993. *The CELEX lexical data base on CD-ROM*. Linguistic Data Consortium, Philadelphia, PA.

Marco Baroni, Johannes Matiasek, and Harald Trost. 2002. Predicting the components of German nominal compounds. In *Proceedings of the 15th European Conference on Artificial Intelligence*. IOS Press, pages 470–474.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods* 46(3):904–911.

Guy Emerson and Thierry Declerck. 2014. SentiMerge: Combining sentiment lexicons in a Bayesian framework. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*.

Andrew Gargett, Josef Ruppenhofer, and John Barnden. 2014. Dimensions of metaphorical meaning. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (COGALEX)*. pages 166–173.

Michael Haas and Yannick Versley. 2015. Subsentential sentiment on a shoestring: A crosslingual analysis of compositional classification. In *HLT-NAACL*. pages 694–704.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. pages 9–15.

Julian Heister, Kay-Michael Würzner, Johannes Bubenzer, Edmund Pohl, Thomas Hanneforth, Alexander Geyken, and Reinhold Kliegl. 2011. dlexDB–eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau* .

Nathalie Japkowicz and Mohak Shah. 2011. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.

Philipp Kanske and Sonja A Kotz. 2010. Leipzig affective norms for German: A reliability study. *Behavior research methods* 42(4):987–991.

Manfred Klenner, Angela Fahrni, and Stefanos Petrakis. 2009. Polart: A robust tool for sentiment analysis. In *Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA 2009)*. pages 235–238.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350000 german lemmas. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. pages 2595–2598.

Olaf Lahl, Anja S. Göritz, Reinhard Pietrowsky, and Jessica Rosenberg. 2009. Using the World-Wide Web to obtain large-scale word norms: 190,212 ratings on a set of 2,654 German nouns. *Behavior Research Methods* 41(1):13–19. https://doi.org/10.3758/BRM.41.1.13.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11):39–41.

Karo Moilanen and Stephen Pulman. 2008. The Good, the Bad, and the Unknown: Morphosyllabic Sentiment Tagging of Unseen Words. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT-Short '08, pages 109–112. http://dl.acm.org/citation.cfm?id=1557690.1557719.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2011. SentiFul: A lexicon for sentiment analysis. *IEEE Transactions on Affective Computing* 2(1):22–36.

Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. 1957. The measurement of meaning. *Urbana: University of Illinois Press* .

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Josef Ruppenhofer, Michael Wiegand, and Jasper Brandes. 2014. Comparing methods for deriving intensity scores for adjectives. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*. Association for Computational Linguistics, Gothenburg, Sweden, pages 117–122. http://www.aclweb.org/anthology/E14-4023.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pages 1–15.

Nathan Schneider, Dirk Hovy Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM). *Proceedings of SemEval* pages 546–559.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*. Citeseer, volume 1631, page 1642.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '11, pages 680–690. http://dl.acm.org/citation.cfm?id=2145432.2145511.

Ulli Waltinger. 2010. GermanPolarityClues: A Lexical Resource for German Sentiment Analysis . In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association, pages 1638–1642.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics* 30(3):277–308.

Michael Wiegand, Christine Bocionek, and Josef Ruppenhofer. 2016. Opinion Holder and Target Extraction on Opinion Compounds - A Linguistic Approach. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. The Association for Computational Linguistics, pages 800–810. http://aclweb.org/anthology/N/N16/N16-1094.pdf.