Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5010–5016
Marseille, 11–16 May 2020

Enhancing a Lexicon of Polarity Shifters through the Supervised Classification of Shifting Directions

Marc Schulder[*†], **Michael Wiegand**[*‡] **and Josef Ruppenhofer**[○]
* Spoken Language Systems, Saarland Informatics Campus, Saarland University, Germany
† Institute for German Sign Language, University of Hamburg, Germany
‡ Leibniz ScienceCampus, Heidelberg/Mannheim, Germany
○ Institute for German Language, Mannheim, Germany
marc.schulder@uni-hamburg.de, wiegand@ids-mannheim.de, ruppenhofer@ids-mannheim.de

Abstract

The sentiment polarity of an expression (whether it is perceived as positive, negative or neutral) can be influenced by a number of phenomena, foremost among them negation. Apart from closed-class negation words like *no*, *not* or *without*, negation can also be caused by so-called polarity shifters. These are content words, such as verbs, nouns or adjectives, that shift polarities in their opposite direction, e. g. *abandoned* in *"abandoned hope"* or *alleviate* in *"alleviate pain"*. Many polarity shifters can affect both positive and negative polar expressions, shifting them towards the opposing polarity. However, other shifters are restricted to a single shifting direction. *Recoup* shifts negative to positive in *"recoup your losses"*, but does not affect the positive polarity of *fortune* in *"recoup a fortune"*. Existing polarity shifter lexica only specify whether a word can, in general, cause shifting, but they do not specify when this is limited to one shifting direction. To address this issue we introduce a supervised classifier that determines the shifting direction of shifters. This classifier uses both resource-driven features, such as *WordNet* relations, and data-driven features like in-context polarity conflicts. Using this classifier we enhance the largest available polarity shifter lexicon.

Keywords: Sentiment Analysis; Polarity Shifter; Negation; Lexical Semantics; Lexicon; Supervised Classification

1. Introduction

Polarity shifters are content words that have semantic properties similar to those of negation words like *not*, *no* or *without*. Like negation words, they can change the sentiment polarity of an expression (also known as valence) by either inverting it or reducing its intensity. For example, the negated statement in (1) can also be expressed by the verbal polarity shifter *fail* or its nominal and adjectival equivalents, as seen in (2)–(4).

(1) Peter [did **not**$_{\textbf{\textit{negation}}}$ [pass the exam]$^{+}$]$^{-}$.
(2) Peter [**failed**$_{\textbf{\textit{shifter}}}$ to [pass the exam]$^{+}$]$^{-}$.
(3) Peter's [**failure**$_{\textbf{\textit{shifter}}}$ to [pass the exam]$^{+}$]$^{-}$
(4) Peter's [**failed**$_{\textbf{\textit{shifter}}}$ [attempt to pass the exam]$^{+}$]$^{-}$

Many polarity shifters can affect both positive and negative expressions. In (5), the verbal shifter *destroy* shifts a positive polar expression to negative, while in (6) it shifts from negative to positive.

(5) It [**destroyed**$_{\textbf{\textit{shifter}}}$ their [hopes]$^{+}$]$^{-}$.
(6) The medication will [**destroy**$_{\textbf{\textit{shifter}}}$ the [cancer]$^{-}$]$^{+}$.

Other shifters, however, are unidirectional and only shift polarities in one direction (Wilson et al., 2005). The verbal shifter *to risk*, for example, shifts only positive polar expressions like *good health* in (7), while the polarity of negative polar expressions like *war* in (8) remains unaffected. Similarly, the adjectival shifter *antiquated* shifts the positive noun *ideal* in (9), but not the negative noun *stereotype* in (8).

(7) You [**risk**$_{\textbf{\textit{shifter}}}$ your [good health]$^{+}$]$^{-}$.
(8) Their actions [**risk** a [war]$^{-}$]$^{-}$.
(9) The "American dream" is an [**antiquated**$_{\textbf{\textit{shifter}}}$ [ideal]$^{+}$]$^{-}$.
(10) Women belonging in the kitchen is an [**antiquated** [stereotype]$^{-}$]$^{-}$.

Conversely there are shifters that only affect negative expressions but not positive ones, such as *recoup* in (11) and (12) and *amend* in (13) and (14).

(11) She must [**recoup**$_{\textbf{\textit{shifter}}}$ her [losses]$^{-}$]$^{+}$.
(12) I could [**recoup** a [fortune]$^{+}$]$^{+}$.
(13) Let us [**amend**$_{\textbf{\textit{shifter}}}$ that [problem]$^{-}$]$^{+}$.
(14) We can [**amend** the [solution]$^{+}$]$^{+}$ to improve its clarity.

Wilson et al. (2005) specify shifting directions by marking the shifter as *'general polarity shifter'*, *'positive polarity shifter'* or *'negative polarity shifter'*, where *positive* and *negative* refer to the polarity that the shifted expression receives. We found that in practice this terminology could cause confusion as to whether the prior or shifted polarity was being referred to. It is also unclear how to use it in cases where shifting results in a neutral polarity expression (Taboada et al., 2011). We therefore choose to instead mark shifters as *'affects positive polarity'*, *'affects negative polarity'* or *'affects both polarities'*.

While multiple resources exist that identify polarity shifters for English (Schulder et al., 2017; Schulder et al., 2018b; Schulder et al., forthcoming) and German (Schulder et al., 2018a), none of them specify their shifting direction. As

a result, the polarities of sentences such as (8), (10), (12) and (14) would erroneously be assumed to have shifted.
To prevent such mistakes, **we introduce a supervised classifier for shifting directions** that can enhance available shifter lexica. It labels each shifter as exactly one of three types: shifters that only affect positive polarities, only negative ones or shifters that can affect both.
Our contributions are the following:

1. We design a supervised classifier of shifting directions which uses a combination of features specifically designed for the task as well as features extracted from general-purpose semantic resources.
2. We create a gold standard dataset of 540 words labeled for shifting directions to allow classifier training.
3. We use our classifier to provide shifting direction labels for all shifters found in the largest shifter lexicon available (Schulder et al., forthcoming).

The resulting **lexicon of shifting directions** for English shifters is made **publicly available**.[1]

2. Related Work

The concept of *polarity shifting* was first introduced to natural language processing by Polanyi and Zaenen (2006). Shifting occurs when the sentiment polarity (also referred to as *valence*) of an expression is moved towards the opposite polarity. Shifting can be caused by a number of linguistic phenomena. While Polanyi and Zaenen (2006) took a broad view on this, including the effects of discourse structure, text genre and socio-cultural factors, subsequent research has narrowed the meaning of shifting to polarity changes caused by lexical items. Schulder et al. (2017) further restrict the term to refer only to content words, such as verbs, nouns and adjectives, in contrast to common syntactic negation words, such as *no*, *not* or *without*, which form a closed class.
Wilson et al. (2005) introduced a polarity classifier that takes into account that the lexical polarity of words is affected by a number of contextual phenomena, polarity shifters among them. They also observed that some shifters would only affect specific polarities and took this into account.
While individual negation words are more frequent than individual polarity shifters, Schulder et al. (2018b) showed that overall shifters are at least as frequent, even when only considering verbal shifters. However, so far research that concerns itself with compositional polarity has mostly focussed on negation words (see the survey by Wiegand et al. (2010)). This is at least in part due to the lack of resources that would help identify polarity shifters. While Socher et al. (2013) showed that negation words can be learned implicitly from labeled data, this fails for polarity shifters due to the relative low frequency of individual shifter words compared to negation words (Schulder et al., 2017). This is a general problem for implicit negation learning, even for current state of the art classifiers (Schulder et al., forthcoming).

[1] https://doi.org/10.5281/zenodo.3545947

Dataset	Verbs	Nouns	Adj.	Total
Gold Standard	304	107	129	540
Bootstrapped	676	793	512	1,981
Total	980	900	641	2,521

Table 1: Number of polarity shifters in the English shifter lexicon (Schulder et al., forthcoming), grouped by part of speech. Both the gold standard and the automatically bootstrapped data were verified by an expert annotator.

As implicit learning of shifters is not viable, the need for explicit shifter resources remains. To meet this need, Schulder et al. (2017) introduced the first lexicon of polarity shifters, covering a large number of English verbs. The lexicon was created using a bootstrapping approach that combined automatic classification with human verification to bootstrap a lexicon of almost a thousand shifters while keeping annotation cost low.
Schulder et al. (2018a) extended this approach to cross-lingual bootstrapping, using it to create a lexicon of German verbal polarity shifters. A second lexicon of English verbal shifters was created by Schulder et al. (2018b), this time completely manually to allow the inclusion of more detailed information, such as shifter labels for individual word senses and information regarding which syntactic arguments are affected by specific shifters. While these shifter lexicons were all limited to verbs, Schulder et al. (forthcoming) expand their bootstrapping approach from Schulder et al. (2017) to also cover nouns and adjectives. This increases the size of the bootstrapped lexicon to over 2,500 polarity shifters. None of the available lexicons address the question of shifting direction, however. We therefore introduce a lexicon extension that adds this information through supervised classification.

3. Data

To determine shifting directions through supervised classification, we require specific resources. In section 3.1 we describe the lexicon of polarity shifters that we will enhance with shifting directions. We use part of this lexicon to create a shifting directions gold standard in section 3.2 on which our classifier will be trained and tested. Our data-driven features also require a text corpus, which is described in section 3.3.

3.1. Lexicon of Polarity Shifters

To determine the shifting direction of polarity shifters, we first need to know which words can generally act as shifters. We use the largest polarity shifter lexicon available, the English shifter lexicon we created previously (Schulder et al., forthcoming), which identifies 2,521 shifters among the *WordNet* vocabulary of verbs, nouns and adjectives.[2] In that work we annotated a gold standard of 2,000 randomly sampled words for each of the three parts of speech, respectively. Each word was labeled for whether it was a shifter or not.

[2] https://doi.org/10.5281/zenodo.3365601

	Verbs		Nouns		Adjectives	
Affects	#	%	#	%	#	%
Positive	47	15.5	36	33.6	85	65.9
Negative	86	28.3	15	14.0	8	6.2
Both	171	56.3	56	52.3	36	27.9
Total	304		107		129	

Table 2: Distribution of shifting directions per part of speech among the 540 polarity shifters from the shifter gold standard by Schulder et al. (forthcoming) (see Table 1).

Using a bootstrap classifier trained on this gold standard, they filtered the remaining *WordNet* vocabulary for words that were considered likely to be shifters and had a human annotator verify them. As only about 5–15% of words are shifters (depending on part of speech), this allowed them to reduce the annotation effort significantly, as the majority of words could be discarded before human verification. The resulting distribution of shifters across parts of speech and between the initial gold standard and the verified bootstrapping output is shown in Table 1.

The shifter lexicon only specifies whether words can, in general, cause shifting, but provides no information on their shifting direction. It must also be noted that the aforementioned bootstrapping approach is not applicable to the labelling of shifting directions. When creating a list of shifters, high precision is far more important for the *'shifter'* label than for the *'non-shifter'* label, as it is more acceptable to mark a few shifters as non-shifters than to incorrectly include non-shifters in the list of shifters.

When labelling shifting directions, no such prioritisation can be made, as the three possible directional behaviours of shifters (affecting positive, negative or both polarities) are all equally relevant. To verify the labels would mean having to manually check every single one of them, making automatic classification superfluous. Instead we will introduce a manually annotated gold standard in section 3.2 and then automatically label the remaining words in section 5.2

3.2. Gold Standard

To create a gold standard for our new task of determining shifting direction we label the 540 shifters that were also part of the shifter lexicon gold standard. This gold standard is used to train and evaluate our supervised classifier, with which we will label the remaining 1,981 shifters.

Each word is given exactly one of three labels: *'affects positive polarities'* if a word can only shift from positive to negative, *'affects negative polarities'* if it can only shift from negative to positive, or *'affects both polarities'* if it can cause shifting in either direction.

All annotations are performed by an expert annotator with experience in linguistics and the annotation of polarity shifters. To judge inter-annotator agreement, one of the authors also labeled 200 words independently, resulting in a Cohen's kappa inter-annotator agreement (Cohen, 1960) of $\kappa = 0.65$, indicating substantial agreement.

The resulting label distribution can be seen in Table 2. Interestingly, the individual parts of speech show distinctly different distributions. About half the verbs and nouns are bidirectional, but among adjectives only a quarter are bidirectional, while two thirds affect only positive words and almost none affect only negative words.

3.3. Text Corpus

Apart from gold standard training data, several of our features also require a textual corpus to perform pattern recognition and to compare word frequencies. We use *Amazon Product Review Data* (Jindal and Liu, 2008), a corpus of 5.8 million product reviews, as it has previously been shown to be a good fit for polarity shifter classification (Schulder et al., 2017).

To prepare the corpus for use in our features, we lemmatise it and merge particle verbs to be represented as a single token (e. g. `tear_down`). To determine syntactic dependency relations within the corpus, we parse it using the *Stanford Parser* (Chen and Manning, 2014).

4. Methodology

To automatically classify shifting directions, we train an SVM multi-class classifier, using the `SVM`$^{\texttt{multiclass}}$ implementation by Tsochantaridis et al. (2005). We train the classifier using features that utilise existing linguistic resources as well as patterns in textual data. In addition we present several baselines.

4.1. Baselines

We define two majority classifiers and a word embedding classifier as baselines. All baselines were also tested for their use as classifier features during exploratory experiments, but as each resulted in decreased performance for the best classifier, we evaluate them only as stand-alone baseline classifiers at this point.

Majority: BASELINE$_{maj}$ assigns the overall majority label *'affects both'* to all words, based on the directionality distribution observed in our gold standard (Table 2).

POS-specific majority: BASELINE$_{pos_maj}$ assigns to each word the majority label for its respective part of speech, i. e. verbs and nouns are still labeled as *'affects both'*, but adjectives receive the label *'affects positives'*. This is a stronger baseline than BASELINE$_{maj}$, as it takes into account the label distributions of individual parts of speech as observed in Table 2.

Word Embedding: For BASELINE$_{embed}$ we train an SVM classifier on the dimensions of a word embedding.[3] Word embeddings are vector spaces that represent semantic similarity based on distributional similarity.

As embedding we use the 500-dimensional `Word2Vec` (Mikolov et al., 2013) embedding of the *Amazon Product*

[3]Classifiers using contextualised embeddings, e. g. `BERT` (Devlin et al., 2019), present no advantage for our task, as lexical classification involves no context. In addition, the small number of 540 training items precludes the use of most other deep learning classifiers.

Review Data corpus created by Schulder et al. (2017).[4] Each dimension is treated as a weighted binary feature. Our expectation is that as shifting directionality is a semantic phenomenon, it may be encoded in specific embedding dimensions. Shifters that share shifting direction would be expected to be close to each other on those dimensions.

4.2. Task-specific Features

The following features are designed specifically for their use in the classification of polarity shifters.

Argument Polarity: Many unidirectional shifters are far more frequently used in contexts that involve the polarity that they affect than those with the unaffected polarity. For example, the verbal shifter *"fend off"*, which affects only negative expressions, occurs almost five times as often with negative expressions than with positive ones. Similarly, the verb *"spoil"* occurs almost three times as often with (affected) positive expressions as with (unaffected) negative expressions.

For our feature we count how often the argument of a shifter has a positive or negative polarity in our text corpus, relative to its overall frequency. To determine the polarity of the arguments we use *Subjectivity Lexicon* (Wilson et al., 2005). The argument of the shifter is defined as one of the following dependency relations:

- the direct object of a verb, e. g. *"it [ruined* $_{shifter}$ ***[our hopes]*** $^{+}$ *]* $^{-}$*"*,
- the compound modifier of a noun compound, e. g. *"[**[cancer]** $^{-}$ *cure* $_{shifter}$*]* $^{+}$*"*,
- the prepositional object of a noun with the preposition *of*, e. g. *"The [destruction* $_{shifter}$ *of* ***[my dreams]*** $^{+}$ *]* $^{-}$*"*,
- the modified noun of an attributive adjective, e. g. *"The [exonerated* $_{shifter}$ ***[convict]*** $^{-}$ *]* $^{+}$*"*,
- or the subject of a predicative adjective, e. g. *"The [**[hero]** $^{+}$ *is dead* $_{shifter}$*]* $^{-}$*"*.

This definition of shifter arguments is a simplified representation designed to fit the needs of data-driven features. For more detailed discussions that also address less frequent kinds of shifted arguments, see Wiegand et al. (2018) and Schulder et al. (2018b).

Verb Particles: Particle verbs are verbs that combine with an adverbial particle, such as *"fend off"* in (15) or *"mess up"* in (17). Certain particles have been shown to indicate a particular aspectual property, such as the complete transition to an end state (Brinton, 1985). For example, the particle *out* in *"dry (something) out"* indicates that *"something is dried completely"*.

Schulder et al. (2017) showed that this property of particles can be leveraged to identify shifters, as the transition to a new negative end state of removal or diminishment is inherent in polarity shifting.

We hypothesise that the specific choice of particle can also provide information on the directionality of the shifter. For example, the particle *off* often appears in shifters that only affect negative expressions, such as *"fend off"* in (15) and *"cast off"* in (16), but not in particles that exclusively affect positive expressions. The particle *up*, on the other hand, is often seen in shifters that exclusively affect positive expressions, such as *"mess up"* in (17) and *"dry up"* in (18), but rarely in shifters that only affect negatives.

(15) They [**fended off**$_{\textbf{\textit{shifter}}}$ an [invasion]$^{-}$]$^{+}$.

(16) She [**cast off**$_{\textbf{\textit{shifter}}}$ her [shackles]$^{-}$]$^{+}$.

(17) We [**messed up**$_{\textbf{\textit{shifter}}}$ our [chances of victory]$^{+}$]$^{-}$.

(18) [[Support]$^{+}$ for the campaign has **dried up**$_{\textbf{\textit{shifter}}}$]$^{-}$.

Our feature identifies which verbs have particles and mark the specific particle, to allow the classifier to differentiate between them.

EffectWordNet: In +/−effect theory, events are described as having beneficial (+effect) or harmful effects (−effect) on their objects (Deng et al., 2013; Choi et al., 2014; Choi and Wiebe, 2014). Schulder et al. (2017) showed that −effects are closely related to polarity shifting, as the harmfulness they describe is often caused by the removal, destruction or reduction that is a defining aspect of shifters.

However, while related, −effect and shifters are not identical. Inspecting the verbal shifter lexicon of Schulder et al. (2017), we find that among the shifters that *EffectWordNet* unambiguously identifies as +effect or −effect, 10% are +effects. We expect that these are unidirectional shifters that only affect negative expressions, so that their effect will be perceived as beneficials, such as *"atone"* in (19) and *"improve"* in (20). Encountering shifters that are labeled *'+effect'* could therefore be an indicator for negative-affecting shifters.

(19) He has [**atoned**$_{\textbf{\textit{shifter}}}$ for his [sins]$^{-}$]$^{+}$.

(20) Her [[illness]$^{-}$ has **improved**$_{\textbf{\textit{shifter}}}$]$^{+}$.

To determine the +/−effect of shifters for our feature, we use *EffectWordNet* (Choi and Wiebe, 2014), which provides +/−effect labels for individual word senses of verbs as defined in *WordNet* (Miller et al., 1990). About 1,600 are manually annotated (roughly half of which are labeled +effect or −effect), the rest are labeled automatically via supervised classification.[5] As we label words, rather than word senses, we label a word as +effect or −effect if at least one of its senses is labeled as such. In cases where this would result in a word being labeled as both +effect and −effect, we instead discard the label. As *EffectWordNet* only covers verbs, so does our feature.

[4] The word embedding can be found at `https://doi.org/10.5281/zenodo.3370051`

[5] It might be that some of the +effect labels we encounter among shifters are due to mistakes in this automatic classification. For example, *"atone"* might be argued to be −effect, as atoning would be bad for the sin. Either way, we must rely on the information provided by *EffectWordNet*, as it is the only sizeable source of +/−effect labels.

4.3. Generic Features

These features have been shown to be of general use to a variety of semantic classification tasks.

WordNet: *WordNet* (Miller et al., 1990) is the largest available ontology for the English language. It has been used successfully in a number of different sentiment analysis applications (Esuli and Sebastiani, 2005; Breck et al., 2007; Gyamfi et al., 2009; Choi and Wiebe, 2014; Kang et al., 2014; Flekova and Gurevych, 2016).

WordNet provides a variety of **semantic relations** between words, such as *hypernymy*, *antonymy*, *entailment* and *derivational relatedness*. Words also receive short sense definitions, called *glosses*, **example sentences**, and coarse semantic categories called **lexicographer senses**.

For the supervised classification of polarity shifters, Schulder et al. (2017) use the hypernymy relation, glosses, and lexicographer senses as features. They report these to be the strongest features for their task, especially glosses, which allow the classifier to detect similarities in how semantically similar words are described.

For our more fine-grained classification task of differentiating shifting direction, we expect to require additional ways to distinguish semantic nuances. Apart from hypernymy we also include all other semantic relations available in *WordNet*, as well as example sentences, where available.

While *WordNet* organises words in sets of synonyms (*synsets*) to differentiate between their different word senses, our classification is performed on the word-level. Our features represent each word as the union of synsets that the word belongs to. Glosses and example sentences are encoded as bag of words features. Semantic relations are given as the synset(s) to which the relation directly connects a word. Lexicographer senses are used directly as feature values.

FrameNet: *FrameNet* (Baker et al., 1998) is a semantic resource based on the theory of frame semantics (Fillmore, 1967). It provides semantic frames that identify words with similar semantic behaviour and describes the semantic roles with which other words can interact with them.

FrameNet has been used for a variety of sentiment-related tasks, such as opinion holder and target extraction (Kim and Hovy, 2006), opinion spam analysis (Kim et al., 2015) and stance classification (Hasan and Ng, 2013). Schulder et al. (2017) successfully used frame membership as features for shifter classification under the hypothesis that shifters will cluster in certain frames, such as AVOIDING, which consists exclusively of shifters. We extend this hypothesis by positing that words may not only cluster to certain frames based on whether they are shifters or not, but also that shifters with a certain shifting direction will likely gather in the same frames.

While *FrameNet* contains over 1,200 frames, this still covers only 31% of the gold standard vocabulary. Coverage can be extended by using the semantic-parser `SemaFor` (Das et al., 2010) to infer missing frames (Das and Smith, 2011). Our feature uses this extended form of *FrameNet* to mark the frame membership of individual shifters.

Classifier	Acc.	Prec.	Recall	F1
BASELINE$_{maj}$	48.7	16.2	33.3	21.8
BASELINE$_{pos_maj}$	57.8†	40.5†	45.7†	42.9†
BASELINE$_{embed}$	60.7	58.1†	58.1†	58.1†
SVM$_{task\text{-}specific}$	48.9	46.3	37.0	40.6
SVM$_{generic}$	67.0*†	69.3*†	61.3†	65.0*†
SVM$_{task+generic}$	**69.6***†	**72.5***	**64.8***†	**68.4***†

*: is better than all baselines (paired t-test with $p < 0.05$)
†: is better than previous classifier (paired t-test with $p < 0.05$)

Table 3: Results of the shifting direction classification. The evaluation is run as a 10-fold cross validation. F-score, precision and recall are macro-averages. Best results are depicted in bold.

5. Experiments

In this section we evaluate the performance of our classifier in section 5.1 and then use the classifier to label the remaining shifters that are not part of our gold standard in section 5.2

5.1. Classifier Evaluation

We treat the labelling of shifting directions as a classification task for polarity shifters of three disjunct classes. Each shifter must be classified as exactly one of three labels: *'affects positive'*, *'affects negative'* and *'affects both'* (see section 3.2).

To evaluate our shifting direction classifier, we perform 10-fold cross validation, repeatedly training the classifier on 90% of the gold standard data and testing it on 10% until all tokens have been part of the test set once. Results are given as the mean value of the 10 repetitions. We report F-score, precision, recall and accuracy. In the case of F-score, precision and recall, we report the macro-averages across the three labels.

Table 3 shows the performance of the classifiers when trained on either the task-specific features, the generic features or on both feature sets together and compares it to our baselines.

BASELINE$_{embed}$ represents a strong supervised baseline. It clearly outperforms the two majority label baselines[6] and shows that the semantic information provided by a word embedding is a solid feature for classifying shifting directions. Unfortunately, we found that using the word embedding as a feature in our multi-feature SVM classifier did not integrate well with the other features.

Looking at SVM$_{task\text{-}specific}$, we see that the task-specific features on their own are not sufficient. This is in part due to the fact that only one of the three features, argument polarity, is available for all parts of speech, while the other two can only be applied to verbs. Even when evaluating only on verbs, though, the classifier cannot beat BASELINE$_{embed}$.

[6]The high accuracy results of the two majority baselines are due to the majority label bias inherent in the metric, rather than actual strong performance.

	Verbs		Nouns		Adjectives	
Affects	#	%	#	%	#	%
Positive	66	9.8	158	19.9	471	92.0
Negative	154	22.8	20	2.5	5	1.0
Both	456	67.5	615	77.6	36	7.0
Total	676		793		512	

Table 4: Distribution of shifting directions among the 1,981 shifters that were automatically labeled (cf. gold standard labels in Table 2). Labels were determined using the best shifting direction classifier from Table 3.

$\text{SVM}_{generic}$ fares considerably better, clearly outperforming all baselines. The task-specific features are not pointless, however, as combining them with the generic features in $\text{SVM}_{task+generic}$ improves performance further, resulting in the strongest available classifier, whose F-score is 10 points above that of the word embedding baseline.

5.2. Extension of the Shifter Lexicon

We use our best classifier, which contains all task-specific and generic features, to classify the remaining 1,981 polarity shifters that were not part of our gold standard. This covers the 676 verbs, 793 nouns and 512 adjectives that were bootstrapped for the shifter lexicon (see Table 1).

Table 4 shows the distribution of shifting direction labels among the automatically classified shifters. We see the same trends as for our gold standard (see Table 2), albeit with a stronger bias towards the majority label of each part of speech.

Combining the shifting direction labels from our gold standard with those from our automatic classification provides us with direction labels for all 2,521 shifters of the polarity shifter lexicon. This lexicon extension is made publicly available (see footnote 1).

6. Conclusion

In this paper we addressed the task of determining the shifting directions of polarity shifters, i. e. whether they affect expressions of any sentiment polarity or only those of either positive or negative polarity.

We developed a new gold standard for this task and found that while many shifters affect both polarities, a significant number of them can only shift in one direction, especially among adjectival polarity shifters. Building on our gold standard, we designed a supervised classifier that uses features derived from general-purpose semantic resources as well as data-driven features designed specifically for determining shifting directions. Using this classifier and our gold standard, we determined shifting direction labels for the complete polarity shifter lexicon of Schulder et al. (forthcoming). The resulting lexicon extension is publicly available.

7. Acknowledgements

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

8. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 86–90, Vancouver, British Columbia, Canada. ACL.

Breck, E., Choi, Y., and Cardie, C. (2007). Identifying Expressions of Opinion in Context. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2683–2688, Hyderabad, India, January.

Brinton, L. J. (1985). Verb Particles in English: Aspect or Aktionsart. *Studia Linguistica*, 39(2):157–168, December.

Chen, D. and Manning, C. D. (2014). A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. ACL.

Choi, Y. and Wiebe, J. (2014). +/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191, Doha, Qatar. ACL.

Choi, Y., Deng, L., and Wiebe, J. (2014). Lexical Acquisition for Opinion Inference: A Sense-Level Lexicon of Benefactive and Malefactive Events. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, pages 107–112, Baltimore, Maryland, USA. ACL.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Das, D. and Smith, N. A. (2011). Semi-Supervised Frame-Semantic Parsing for Unknown Predicates. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT)*, pages 1435–1444, Portland, Oregon, USA. ACL.

Das, D., Schneider, N., Chen, D., and Smith, N. A. (2010). Probabilistic Frame-Semantic Parsing. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 948–956, Los Angeles, California, USA. ACL.

Deng, L., Choi, Y., and Wiebe, J. (2013). Benefactive/Malefactive Event and Writer Attitude Annotation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 120–125, Sofia, Bulgaria. ACL.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, Minneapolis, Minnesota. ACL.

Esuli, A. and Sebastiani, F. (2005). Determining the Semantic Orientation of Terms through Gloss Classifica-

tion. In *Proceedings of the Conference on Information and Knowledge Management (CIKM)*, pages 617–624, Bremen, Germany. ACM.

Fillmore, C. J. (1967). The Case for Case. In Emmon Bach et al., editors, *Proceedings of the Texas Symposium on Language Universals*, pages 1–90, New York City, New York, USA, April. Holt, Rinehart and Winston.

Flekova, L. and Gurevych, I. (2016). Supersense Embeddings: A Unified Model for Supersense Interpretation, Prediction, Utilization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2029–2041, Berlin, Germany. ACL.

Gyamfi, Y., Wiebe, J., Mihalcea, R., and Akkaya, C. (2009). Integrating Knowledge for Subjectivity Sense Labeling. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 10–18, Boulder, Colorado, USA. ACL.

Hasan, K. S. and Ng, V. (2013). Frame Semantics for Stance Classification. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 124–132, Sofia, Bulgaria. ACL.

Jindal, N. and Liu, B. (2008). Opinion Spam and Analysis. In *Proceedings of the International Conference on Web Search and Data Mining (WSDM)*, pages 219–230, Palo Alto, California, USA. ACM.

Kang, J. S., Feng, S., Akoglu, L., and Choi, Y. (2014). ConnotationWordNet: Learning Connotation over the Word+Sense Network. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1544–1554, Baltimore, Maryland. ACL.

Kim, S.-M. and Hovy, E. (2006). Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text (SST@ACL)*, pages 1–8, Sydney, Australia. ACL.

Kim, S., Chang, H., Lee, S., Yu, M., and Kang, J. (2015). Deep Semantic Frame-Based Deceptive Opinion Spam Analysis. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 1131–1140, Melbourne, Australia, October. ACM.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the Workshop at the International Conference on Learning Representations (ICLR)*, Scottsdale, Arizona, USA.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3:235–244.

Polanyi, L. and Zaenen, A. (2006). Contextual Valence Shifters. In James G. Shanahan, et al., editors, *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, pages 1–10. Springer Netherlands, Dordrecht, Netherlands.

Schulder, M., Wiegand, M., Ruppenhofer, J., and Roth, B. (2017). Towards Bootstrapping a Polarity Shifter Lexicon using Linguistic Features. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 624–633, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Schulder, M., Wiegand, M., and Ruppenhofer, J. (2018a). Automatically Creating a Lexicon of Verbal Polarity Shifters: Mono- and Cross-lingual Methods for German. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 2516–2528, Santa Fe, New Mexico, USA, August. ICCL.

Schulder, M., Wiegand, M., Ruppenhofer, J., and Köser, S. (2018b). Introducing a Lexicon of Verbal Polarity Shifters for English. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1393–1397, Miyazaki, Japan, May. ELRA.

Schulder, M., Wiegand, M., and Ruppenhofer, J. (forthcoming). Bootstrapped Creation of a Lexicon of Sentiment Polarity Shifters. *Journal of Natural Language Engineering*, (Special Issue on Processing Negation).

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, Seattle, Washington, USA. ACL.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307, March.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005). Large Margin Methods for Structured and Interdependent Output Variables. *Journal of Machine Learning Research (JMLR)*, 6(Sep):1453–1484.

Wiegand, M., Balahur, A., Roth, B., Klakow, D., and Montoyo, A. (2010). A Survey on the Role of Negation in Sentiment Analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP)*, pages 60–68, Uppsala, Sweden.

Wiegand, M., Wolf, M., and Ruppenhofer, J. (2018). Negation Modeling for German Polarity Classification. In Georg Rehm et al., editors, *Language Technologies for the Challenges of the Digital Age*, Lecture Notes in Computer Science, pages 95–111. Springer International Publishing, January.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Joint Conferences on Human Language Technology and on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 347–354, Vancouver, British Columbia, Canada. ACL.