Towards Bootstrapping a Polarity Shifter Lexicon using Linguistic Features

Marc Schulder*, Michael Wiegand*, Josef Ruppenhofer†, Benjamin Roth‡
* Spoken Language Systems, Saarland University
† Institute for German Language, Mannheim
‡ Center for Information and Language Processing, LMU Munich

marc.schulder@lsv.uni-saarland.de
michael.wiegand@lsv.uni-saarland.de
ruppenhofer@ids-mannheim.de
beroth@cis.uni-muenchen.de

Abstract

We present a major step towards the creation of the first high-coverage lexicon of polarity shifters. In this work, we bootstrap a lexicon of verbs by exploiting various linguistic features. Polarity shifters, such as *abandon*, are similar to negations (e.g. *not*) in that they move the polarity of a phrase towards its inverse, as in *abandon all hope*. While there exist lists of negation words, creating comprehensive lists of polarity shifters is far more challenging due to their sheer number. On a sample of manually annotated verbs we examine a variety of linguistic features for this task. Then we build a supervised classifier to increase coverage. We show that this approach drastically reduces the annotation effort while ensuring a high-precision lexicon. We also show that our acquired knowledge of verbal polarity shifters improves phrase-level sentiment analysis.

1 Introduction

We present an approach towards bootstrapping a lexicon of **polarity shifters**. Polarity shifters are content words that have semantic properties similar to negation. For example, the negated statement in (1) involving the negation word *not* can also be expressed by the verbal shifter *fail* in (2).

(1) Peter did **not** pass the exam.
(2) Peter **failed**$_{shifter}$ to pass the exam.

Similarly, shifting is also caused by nouns (e.g. *downfall*) and adjectives (e.g. *devoid*).

Polarity shifters are important for various tasks in NLP, such as relation extraction (Sanchez-Graillet and Poesio, 2007), recognition of textual entailment (Harabagiu et al., 2006) and particularly sentiment analysis (Wiegand et al., 2010).

Similarly to negation words, they may cause the polarity of a statement to shift. Even though (3) contains the positive polar expression *scholarship*, the overall polarity of the sentence is negative. (4) conveys positive polarity despite the presence of the negative polar expression *pain*.

(3) She was [**denied**$_{shifter}$ the [scholarship]$^{+}$]$^{-}$.
(4) The new treatment has [**alleviated**$_{shifter}$ her [pain]$^{-}$]$^{+}$.

Although there has been significant research on polarity shifting in sentiment analysis (Wiegand et al., 2010), this work has focused on the presence of negation words. Negation words (*no*, *not*, *never*, etc.) are typically function words, so only a few exist. Polarity shifters are content words, of which there are a lot more. For instance, WordNet (Miller et al., 1990) contains over 10k verbal, 20K adjectival and 110K nominal lemmas. An exhaustive manual annotation would be far too costly.

To reduce cost, we introduce a bootstrapping approach for the acquisition of polarity shifters. In this work we focus exclusively on verbs. As the main predicates of phrases they tend to have larger scopes than nouns and adjectives, increasing the impact of their polarity shifting. Their vocabulary size is also smaller, allowing us to cover a reasonable share of it in our evaluation.

Existing resources barely cover any verbal shifters at all. Even the most complex negation lexicon for sentiment analysis (Wilson et al., 2005) includes a mere 12 verbal shifters. In contrast, our initial random sample of 2000 verb lemmas contained 300 shifters. The corpora on which negation can be learned, such as the Sentiment Treebank (Socher et al., 2013) or the BioScope corpus (Szarvas et al., 2008), only comprise contiguous sentences of fairly small datasets, so only the most frequently occurring negation words are considered. For example, only 6 verbal shifters are observed on the BioScope corpus (Morante, 2010).

In this work, we address this knowledge gap by bootstrapping a lexicon of polarity shifters. On a sample of manually annotated verbs extracted from WordNet, we first examine a variety of linguistic features for this task. Then we build a supervised classifier to classify the remaining WordNet verbs. Thus we can drastically cut down on the number of verbs to be annotated manually.

Our **contributions** are as follows:

(i) we present the first high-coverage lexicon of verbal polarity shifters, going substantially beyond what can be extracted from existing phrase-level corpus annotations;

(ii) we develop methods for high-precision recognition of polarity shifters;

(iii) in addition to using resource-based generic features, we show that we can boost performance with novel task-specific features, many of which are derived from corpora; and

(iv) we show that compositional classification based on recognition of polarity shifters significantly outperforms polarity classifiers that lack explicit knowledge of verbal shifters.

The main focus of this paper is to find ways to automatically *extract* verbal shifters. The question of their respective scope is part of future research.[1] While our work focuses on English, the concepts applied are not language-specific. We have made all data that were manually labeled as part of this research **publicly available**.[2]

2 Data

To obtain a gold standard for verbal shifters, an expert annotator labeled a random sample of 2000 verbs taken from WordNet 3.1 *(see footnote 2)*. To measure interannotator agreement, 200 of these were annotated again by one of the authors. The achieved $\kappa = 0.66$ indicates substantial agreement (Landis and Koch, 1977).

Due to the lack of robust word-sense disambiguation, our annotation is on the lemma level. We follow a simple **binary classification**: each verb either can cause polarities to shift or not. In order to qualify as a shifter, the verb must allow polar expressions as dependents and the polarity of the proposition that embeds both the verb and

---
[1]We assume that the scope of a verbal shifter is the set of its dependents (typically its subject or objects).

[2]https://github.com/marcschulder/ijcnlp2017

	Freq	Perc
shifter	304	15.20
no shifter	1696	84.80

Table 1: Distribution of verbal shifters in annotated sample of WordNet 3.1.

	Polar Verbs		Positive V.		Negative V.	
	Freq	Perc	Freq	Perc	Freq	Perc
shifter	53	18.8	4	5.5	49	25.9
no shifter	229	81.2	69	94.5	140	74.1

Table 2: Distribution of verbal shifters in the *Subjectivity Lexicon* (Wilson et al., 2005).

polar expression must move towards a polarity that is opposite of the polar expression.

Table 1 shows the distribution of shifters among the set of verbs. At approximately 15%, shifters represent a large enough proportion of verbs to be considered for automatic extraction. Table 2 shows the distribution of shifters among polar verbs. (Polar expressions are identified with the help of the *Subjectivity Lexicon* (Wilson et al., 2005).) While a large number of shifters are themselves negative polar expressions, not all are.

3 Task-Specific Features

In the following we present task-specific features for both verbal shifters (§3.1) and their counterpart, anti-shifters (§3.2). Each feature creates a verb ranking, indicating how likely each verb is to be considered a verbal (anti-)shifter.

Most of our features are corpus-based. As a corpus we use *Amazon Product Review Data* (Jindal and Liu, 2008), comprising over 5.8 million reviews. We chose this dataset for its size and its sentiment-related content. Some features also require knowledge of opinion words and their respective polarity. Such knowledge is obtained from the *Subjectivity Lexicon*. Whenever we use syntactic information (e.g. dependency relations), we obtain it from the Stanford Parser (Chen and Manning, 2014).

3.1 Features for Shifter Detection

As baseline features, we consider all verbs ranked by their frequency in our text corpus (**FREQ**), as well as all negative polar expressions[3] ranked by frequency (**NEGATIVE**).

---
[3]We consider negative polar expressions since the proportion of shifters is greatest among these expressions (Table 2).

EffectWordNet (EFFECT): This feature uses the idea that events may have beneficial or harmful *effects* on their objects. Wiebe and colleagues (Deng et al., 2013; Choi et al., 2014; Choi and Wiebe, 2014) introduced this idea in the context of annotation and lexical acquisition work for opinion inference.[4] For example, in (5) the combined facts that *fall* has a negative effect (henceforth referred to as –*effect*) on its theme (i.e. *Chavez*) and that people are happy about Chavez' fall suggest that people have a negative attitude towards Chavez. As (5) shows, verbs with a –effect, such as *fall*, often coincide with verbal shifters. However, –effect words do not necessarily shift polarity. For instance, while *abuse* has a –effect, it does not shift, as shown by the fact that the verb phrase remains negative in (6).

(5) I think people are happy because $[[\text{Chavez}]^- \text{ has}$ **fallen**$_{-\mathit{effect}}]^+$.

(6) We don't want the public getting the idea that we $[\textbf{abuse}_{-\mathit{effect}} \text{ our } [\text{prisoners}]^-]^-$.

While –effect and shifting are not equivalent, their large overlap warrants investigating. We use the related EffectWordNet resource (Choi and Wiebe, 2014), which provides effect labels for synsets. To generalize to lemmas, we label a word as –effect if at least one of its synsets has a –effect and none have a +effect. Analogous to negative polar expressions, we take all –effect words and sort them by word frequency (−EFFECT).

Distributional Similarity (SIM): As our aim is to identify verbs whose semantics resembles that of negation words, a straightforward method is to extract verbs that are distributionally similar to negation words. Using Word2Vec (Mikolov et al., 2013), we compute word embeddings for our text corpus.[5] All verbs are ranked by their cosine similarity to a given negation word. The highest ranking verbs are considered verbal shifters. As negation words we consider the intersection of two negation word lists: the *negation* category in the valence shifter lexicon by Wilson et al. (2005) and the negation signals from Morante and Daelemans (2009). The negation words are *neither*, *never*, *no*, *none*, *nor*, *not* and *without*.

---

[4]Initially, events with positive/negative effects were referred to as *good-for/bad-for* events. We use the terminology Choi and Wiebe (2014) introduced for EffectWordNet.

[5]Following the work of Wiegand and Ruppenhofer (2015) in verb category induction for sentiment roles, a task similar to ours, we use continuous bag-of-words with 500 dimensions.

Polarity Clash (CLASH): Some of our previous examples (e.g. (3)) suggest that shifting is mainly caused by a polar verb (e.g. *lose*$^-$, *alleviate*$^+$, *deny*$^-$) modifying a polar expression with the opposite polarity (e.g. $[[\mathit{lose}_{\mathit{shifter}}]^- [\mathit{hope}]^+]^-$, $[[\mathit{alleviate}_{\mathit{shifter}}]^+ [\mathit{pain}]^-]^+$, or $[[\mathit{deny}_{\mathit{shifter}}]^- [\mathit{scholarship}]^+]^-$). We expect that the more often a verb occurs within such constructions, the more likely it is to be a shifter. As we saw in the *Subjectivity Lexicon* that the majority of verbal shifters have negative polarity (Table 2), we look exclusively for negative polar verbs that have a positive polar noun as a direct object. We rank those verbs by the frequency of occurring with positive nouns (CLASH), normalized by the overall frequency of the verb (CLASH$_{\text{norm}}$).

Particle Verbs (PRT): With many particle verbs, the particles signal a particular aspectual property, typically the occurrence of a complete transition to an end state (Brinton, 1985). For instance, *dry (something) out* means *dry (something) completely*. Since shifting normally involves producing a new (negative) end state of some entity, we assume a significant number of shifters among particle verbs ((7) and (8)).

(7) This $[\textbf{tore\_down}_{\mathit{shifter}} \text{ our great } [\text{dream}]^+]^-$.

(8) Please $[\textbf{lay\_aside}_{\mathit{shifter}} \text{ all your } [\text{worries}]^-]^+$.

We only consider particles which typically indicate a complete transition to a negative end state: *aside*, *away*, *back*, *down*, *off* and *out*. To produce rankings, we sort the particle verbs by their absolute frequency in our text corpus.

Heuristic using 'any' (ANY): Our final shifter feature rests on the linguistic insight that *negative polarity items (NPIs)* (Giannakidou, 2008), such as English *any*, typically appear in the context of a negation, as in (9). Our assumption is that NPIs may similarly occur in the context of a verbal shifter, as in (10), since it similarly conveys a negation. The concept of NPIs is not specific to the English language and can be found in many other languages (Krifka, 1991).

(9) They did $[\textbf{not} \text{ give us } \underline{\text{any}} [\text{help}_{\mathit{dobj}}]^+]^-$.

(10) They $[\textbf{denied}_{\mathit{shifter}} \text{ us } \underline{\text{any}} [\text{help}_{\mathit{dobj}}]^+]^-$.

The feature we design collects all verbs that take a direct object that is modified by the NPI *any*, as in (10). We sort the verbs by their frequency of co-occurrence with this particular textual pattern (ANY). We normalize that pattern frequency by the frequency of the respective verb

(ANY_{norm}). As a further constraint we demand that the direct object represents a polar expression ($\text{ANY}_{\text{norm+polar}}$). This constraint is fulfilled in (10) since *help* is a positive polar expression.

3.2 Anti-Shifter Feature (ANTI)

We also introduce a feature for automatically retrieving verbs that – semantically speaking – are the exact opposite of what shifters convey. This is, therefore, a negative feature indicating the absence of a shifter. Our anti-shifter feature determines verbs co-occurring with a very small set of specific adverbials. Using the log-likelihood collocation measure of *Sketch Engine*[6] we select adverbials that showed attraction to verbs of creation on the one hand, and being repelled by verbs of destruction on the other. Verbs of creation are expected to be anti-shifters, since they typically entail a positive end state (i.e. something is created), while verbs of destruction typically entail a negative end state (i.e. something is destroyed) We identified four different adverbials: *exclusively*, *first*, *newly* and *specially*. Some typical examples are given in (11)–(14). In order to produce a ranking for this feature, we sort the anti-shifter candidate verbs according to their frequency of co-occurrence with either of the respective adverbs, normalized by the respective verb frequency (ANTI).

(11) In winter, black bears exclusively **live**$_{antiShifter}$ on fish.
(12) Full keyboards on cellphones were first **introduced**$_{antiShifter}$ in 1997.
(13) These buildings have been newly **constructed**$_{antiShifter}$.
(14) They specially **prepared**$_{antiShifter}$ vegan dishes for me.

4 Generic Features

In addition to the task-specific features presented in §3 we examine some generic features derived from common lexical resources. Unlike the features in §3, the generic features do not produce a ranking. Therefore, we will only be able to evaluate them in the context of a supervised classifier.

WordNet (WN): WordNet is the largest English ontology. It is organized in synsets. However, we want to assign categories to words, rather than senses. Due to the lack of robust word-sense disambiguation, we represent a word as the union of synsets containing it.

A common way to harness WordNet for lexicon induction tasks in sentiment analysis is by using its **glosses** (Esuli and Sebastiani, 2005; Gyamfi et al.,

---

[6]http://www.sketchengine.co.uk/

2009; Choi and Wiebe, 2014; Kang et al., 2014). We assume that the explanatory texts of glosses are similar among shifters. We treat glosses as a bag-of-words feature.

We also use WordNet to assign semantic types. Our intuition is that verbal shifters share the same semantic types. We consider two types of information that have been previously found effective for sentiment analysis in general, namely the **hypernyms** of verbs (Breck et al., 2007) and their **supersenses** (Flekova and Gurevych, 2016).

FrameNet (FN): FrameNet (Baker et al., 1998) is a semantic resource used for various sentiment related tasks, such as opinion holder and target extraction (Kim and Hovy, 2006), stance classification (Hasan and Ng, 2013) or opinion spam analysis (Kim et al., 2015). It provides over 1200 semantic frames that comprise words with similar semantic behavior. We use the frame-memberships of a verb as its features, expecting that verbal shifters are grouped in the same frames. For instance, the frame AVOIDING exclusively comprises verbal shifters (e.g. *desist*, *dodge*, *evade*, *shun*, *shirk* etc.).

The latest version of FrameNet (v1.6) covers only 31.4% of verbs from our gold standard. To **extend coverage**, we use the semantic-parser SemaFor (Das et al., 2010), which can infer frames for verbs missing from FrameNet (Das and Smith, 2011). For each missing verb, we have SemaFor label 100 sentences from our corpus and use the frame most frequently assigned. In our exploratory experiments with supervised classification, this expansion caused a significant increase of 6% in F-score (paired t-test, $p < 0.05$).

5 Experiments

We will now experimentally evaluate the features introduced in §3 and §4. In §5.1 we analyse the high-precision potential of individual task-specific features (§3). In §5.2 we run a recall-oriented evaluation of our entire gold standard with classifiers using both task-specific and generic features. Using the best classifier from this evaluation, we bootstrap the remaining WordNet verbs into a larger list of shifters in §5.3. Finally, we evaluate the impact of verbal shifter knowledge on phrase-level sentiment analysis in §5.4.

Feature	Retr.	Prec@n			
		20	50	100	250
FREQ	2000	10.0	18.0	22.0	22.0
NEGATIVE	189	30.0	30.0	29.0	N/A
−EFFECT	175	45.0	44.0	46.0	N/A
SIM_{nor}	1901	15.0	24.0	16.0	18.4
$SIM_{neither}$	1901	20.0	18.0	18.0	21.6
SIM_{none}	1901	25.0	24.0	22.0	21.6
SIM_{not}	1901	25.0	24.0	23.0	23.2
SIM_{never}	1901	20.0	30.0	30.0	32.8
SIM_{no}	1901	35.0	28.0	36.0	28.8
$SIM_{without}$	1901	40.0	36.0	34.0	27.6
$SIM_{centroid}$	1901	45.0	30.0	29.0	27.6
CLASH	107	40.0	52.0	39.0	N/A
$CLASH_{norm}$	107	45.0	46.0	37.0	N/A
PRT	165	60.0	64.0	58.0	N/A
ANY	539	30.0	28.0	29.0	34.0
ANY_{norm}	539	65.0	60.0	53.0	38.8
$ANY_{norm+polar}$	272	75.0	66.0	62.0	41.2
$ANY_{norm+polar+pageR}$	1901	**80.0**	**70.0**	**63.0**	**45.2**

Table 3: Analysis of shifter features (§3.1).

Feature	Retrieved	Prec@n			
		20	50	100	250
FREQ	2000	90.0	82.0	78.0	78.0
POSITIVE	73	90.0	94.0	N/A	N/A
+EFFECT	95	90.0	92.0	N/A	N/A
ANTI	725	**95.0**	**96.0**	**93.0**	**87.4**

Table 4: Analysis of anti-shifter feature (§3.2).

5.1 Analysis of Task-Specific Features

In Table 3 we analyze how useful the task-specific shifter features from §3.1 are as high-precision candidate lists. Each feature produces a ranking, which we evaluate in terms of precision at a certain rank (*Prec@n*). We also state the number of retrieved verbs. Embedding-based methods (e.g. SIM) could theoretically rank all verbs. However, the default configuration of Word2Vec discards every word which occurs less than 5 times, which is why only 1901 verbs are retrieved.

Table 3 shows that filtering verbs by effect (−EFFECT) brings improvements over the FREQ and NEGATIVE baselines. Regarding distributional similarity to negation words (SIM), most negation words perform no better than the baselines. The only notable exception is *without*, which provides gains at high ranks. We also examined a combination of all negation words by merging them in a centroid vector ($SIM_{centroid}$) but got mixed results. Polarity clashes (CLASH) show good performance. Particles (PRT) are the second best feature while ANY is the best feature. Normalization and polarity restriction are effective.

We try to further improve the best ranking

Classifier	Acc	Prec	Rec	F1
$Baseline_{majority}$	84.8	42.4	50.0	45.9
$kNN_{noAntiShifter}$	67.6*	54.9	56.4	55.6*
kNN	71.5*	58.3	59.6	58.9*
$LP_{noAntiShifter}$	79.1*	63.0	56.6	59.6*
LP	80.7*	68.6	56.7	62.0*
$SVM_{task-spec.\ features\ (§3)}$	79.9*	65.5	69.7	67.5*
$SVM_{generic\ features\ (§4)}$	89.0*	79.6	74.4	76.9*
$SVM_{all\ features}$	**89.7***	**80.7**	**77.6**	**79.1***

*: better than previous feature (paired t-test with $p < 0.05$)

Table 5: Evaluation of classification (§5.2) on the 2000 verbs from gold standard (Table 1).

(i.e. $ANY_{norm+polar}$) by applying **personalized PageRank** (Haveliwala, 2002; Agirre and Soroa, 2009). In traditional *PageRank* a ranking of nodes in a graph is produced where the highest ranked nodes are the ones most highly connected. In *personalized PageRank* prior information is added. A biased graph is constructed in which attention is drawn towards particular regions of interest (i.e. sets of nodes). This is achieved by assigning specific re-entrance weights to the individual nodes.[7] In our case, we build a word-similarity graph where our verbs are nodes and edges encode similarities between them. The similarities are computed in the same fashion as our distributional similarity features (SIM) (§3). As prior information, we set the nodes representing the verbs returned by $ANY_{norm+polar}$ with a uniform re-entrance weight probability while all other nodes receive a weight of 0. We consider a standard setting of $\alpha = 0.1$ (Manning et al., 2008, ch. 21.2). The resulting ranking indeed improves performance.

In Table 4 we analyze our anti-shifter feature (§3.2). As baseline we again consider all verbs ranked by frequency (FREQ). Complementary to NEGATIVE and −EFFECT from Table 3, we consider positive polar expressions (POSITIVE) and +effects (+EFFECT). Our anti-shifter feature (ANTI) clearly outperforms the other approaches.

5.2 Classifier Evaluation

In preparation to our bootstrapping task, we perform a recall-oriented evaluation to consider the classification of *all* verbs from our gold standard as opposed to the n-best rankings used in §5.1. We consider two types of classifiers (as well as a simple majority-class baseline): **graph-based classifiers** and **supervised classifiers**. As graph-based

---

[7] A non-uniform distribution causes some preferred nodes of interest to be visited more often during the random walk.

Conf. Rank	1-250	251-500	501-750	751-1043[†]
Precision	92.8	73.2	62.4	33.2

[†]: final interval covers all remaining predicted *shifters*

Table 6: Classification of WordNet verbs that were **not** part of gold standard (§2); verbs are ranked by confidence-score of classifier and evaluated at intervals by precision of *shifter* label.

classifiers, we use one based on label propagation (**LP**), as well as a k-nearest neighbor classifier (**kNN**). **LP and kNN do *not* employ any manually labeled training data**. We use seeds produced by our best task-specific features (§3.1 and §3.2). Then labels are propagated with the help of a word-similarity graph. We use the graph we already employed for our PageRank experiments in §5.1. As shifter seeds we use the top 250 items from $ANY_{norm+polar+pageR}$. We use twice as many seeds for anti-shifters[8] (using ANTI from §3.2) to reflect the general bias towards non-shifter verbs (Table 1). In order to examine whether anti-shifters are actually necessary to get negative seeds of sufficient quality, we also run an alternative setting (*noAntiShifter*) in which the same number of negative seeds is simply extracted from the ranking of frequent verbs. The reasoning behind this is that the proportion of frequent verbs not being shifters is already fairly high, as shown by FREQ in Table 4. For LP, we considered the Adsorption label propagation algorithm as implemented in `junto` (Talukdar et al., 2008). For kNN, we set $k = 10$.

Apart from the graph-based classifiers, we also consider a **supervised classifier**, namely Support Vector Machines (**SVM**) as implemented in SVM^{light} (Joachims, 1999). This classifier uses manually labeled training data, but, unlike LP and kNN, we may combine arbitrary feature sets. We perform 10-fold cross validation and report on accuracy and macro-average precision, recall and F-score. For the task-specific features (§3) we use their most complex configurations from Table 3 (e.g. $SIM_{centroid}$ rather than SIM_{nor} or $SIM_{without}$).

Table 5 shows that among the graph-based classifiers, LP is notably better than kNN. Both classifiers benefit from anti-shifter seeds. Supervised classification outperforms graph-based classification, so using labeled training data is beneficial.

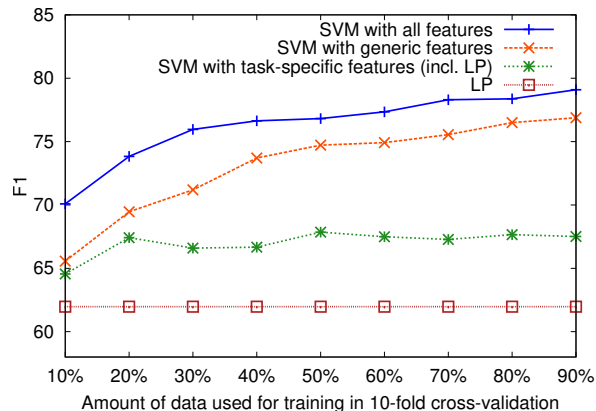[8]This value is a first estimate and could be improved by fine tuning on a development set.

Figure 1: Learning curve on gold standard.

It also means that the *full* set of task-specific shifter features (§3.1) is more effective than just the strongest feature (which is used as seeds for graph-based classification). While the generic features outperform the task-specific features (in supervised classification), combining them results in another significant improvement, demonstrating the importance of the task-specific features.

Figure 1 displays the learning curve of the major feature sets using SVM. While the task-specific features on their own are always worse than the generic features, a classifier combining those feature groups always outperforms the classifier solely trained on the generic features. This improvement is particularly large when few labeled training data are available, which is a typical setting for lexicon bootstrapping tasks. Figure 1 also shows that the SVM classifier has reached roughly the point of saturation when using all features and the maximal amount of labeled training data. This amount should be sufficient for bootstrapping our gold standard lexicon on further unlabeled verbs (as will be shown in §5.3).

5.3 Bootstrapping the Lexicon

We now bootstrap a larger list of shifters from the remaining unlabeled 8581 WordNet verbs not included in our gold standard (§2). On this verb set we run an SVM trained on the gold standard (2000 verbs) with the best performing feature set (Table 5). The classifier predicts 1043 verbs as shifters. The remaining 7538 instances predicted as non-shifters will not be considered further. As our classifier reached a precision of 93.1% on non-shifters on our gold standard data, we are confident that the predicted non-shifters include few actual shifters. As our precision for shifters is lower, i.e.

Information	Example	Label
Sentence	Norah Jones' smooth voice could soothe any savage beast.	
Verb	soothe	
Polar Noun	beast	negative
Verb Phrase	soothe any savage beast	positive

Table 7: Annotation example for the contextual sentiment analysis task. The polarities of the polar noun and the verb phrase are annotated based on context given by the sentence.

Shifting Label	Noun Polarity $\Rightarrow$ VP Polarity		
shifted	$+ \Rightarrow -$	$- \Rightarrow +$	$\sim \Rightarrow +$
	$+ \Rightarrow \sim$	$- \Rightarrow \sim$	$\sim \Rightarrow -$
not shifted	$+ \Rightarrow +$	$- \Rightarrow -$	$\sim \Rightarrow \sim$

Table 8: How the shifting label is derived from the polarities of the polar noun and the verb phrase (positive ($+$); negative ($-$); neutral ($\sim$)).

	Classifier	Acc	Prec	Rec	F1
Baseline	Majority	79.9	39.9	50.0	44.4
	RNTN	59.0	50.8	51.2	51.0
Proposed	LEX_{LP}	84.3	77.7	67.4	72.2
	LEX_{SVM}	87.1	80.0	79.4	79.7
	LEX_{gold}	90.8	88.9	81.2	84.8

Table 9: Evaluation of polarity classification.

68.3%, we manually check the predicted shifter instances. Using our classifier to pre-filter the data (Choi and Wiebe, 2014) reduced the amount to be annotated by 87.8% from 8581 to just 1043 instances. This is an enormous reduction in annotation effort. Table 6 shows the precision on different intervals ranked by confidence score of the SVM on the predicted 1043 shifters. Since the top 250 instances reach a very high precision, with hindsight, a manual annotation of at least these instances would not even have been necessary either.

Among the 1043 predicted shifters, manual annotation confirmed 676 actual shifters. In total we produced a novel list of 980 shifters (304 gold standard + 676 bootstrapping) in this paper *(all included in our public dataset (see footnote 2))*.

5.4 Impact on Sentiment Analysis

We now investigate whether knowledge of verbal shifters can be useful for the identification of contextual **phrase-level sentiment**. Apart from being an intermediate step in compositional sentence-level classification, phrase-level classification is also independently needed for applications such as knowledge base population (Mitchell, 2013), question answering (Dang, 2009) and summarization (Stoyanov and Cardie, 2011). For that reason and because we specifically study compositionality between verbs and their object, we exclusively consider polarity classification for verb phrases.

The experiment is treated as a binary classification task, where the polarity of a noun has either shifted in the context of a verb phrase (VP) or not. For example in (15), the VP *lack her usual passion* contains the positive polarity noun *passion* which is shifted by *lack*.

(15) The book seemed to $[\mathbf{lack}_V$ [her usual $\mathbf{passion}_N^+]_{NP}]_{VP}^-$.

We compiled sentences from our text corpus (*Amazon Product Review Data*, §2) that contain

a VP headed by a verb that has a polar noun[9] as a dependent. We annotated 400 randomly sampled sentences in which the verb is a verbal shifter. We then annotated 2231 sentences with non-shifters to match the ratio of shifters and non-shifters in the gold standard (Table 1) *(see footnote 2)*.

To cover a variety of different verbs, rather than just the most frequent ones, each verbal shifter occurs only once. For each sentence, an annotator labeled the polarity of the polar noun and the polarity of the VP as either *positive*, *negative* or *neutral*. The annotator was also given the full sentence to establish context and the verb that is the head of the VP to avoid misunderstandings. Table 7 shows an example of the information provided, as well as the annotator's decision to label *beast* as negative and *soothe any savage beast* as positive.

Depending on whether the VP and its dependent noun have the same polarity or not, the polarity is considered to have *shifted* or *not shifted*, as detailed in Table 8. These are the class labels onto which the output of the systems (and the annotation) will be mapped. The quantitative evaluation happens on these labels. There is currently no consensus as to how shifting is to be modeled in terms of resulting polarities. For example, the shifting of *excellent* in (16) could either be interpreted as the resulting phrase *wasn't excellent* carrying negative or neutral polarity. The first interpretation simply flips the polarity (Choi and Cardie, 2008), while the second interpretation is driven by the fact that the negation of *excellent* is not synonymous with its antonym *atrocious* (Taboada et al., 2011; Kiritchenko and Mohammad, 2016). The polar in-

---

[9] Noun polarity is provided by the *Subjectivity Lexicon*.

tensity of *wasn't excellent* is certainly weaker than that of *atrocious* but it is a matter of interpretation whether to classify it as negative or neutral. To accommodate both legitimate interpretations, we count either of these behaviors as shifting (Table 8). We do this since our evaluation is concerned with whether shifting occurs, not with the exact polarities (or polar intensities) involved. Our own approach does not profit from this, as it is based on the knowledge of shifters, not polarities.

(16) Let's say, the movie [**wasn't** [excellent]$^+$]$^{-/\sim}$.

As baselines, we consider a majority class classifier (**Majority**) and the Recursive Neural Tensor Network tagger (**RNTN**) by Socher et al. (2013), which is considered the state-of-the-art for handling negation on the phrase level. RNTN is a compositional sentence-level polarity classifier providing polarity values for each tree node in the constituency parse of a sentence. This output allows us to extract polarity predictions for VPs and polar nouns in our data. Apart from achieving best performance on polarity classification datasets, a major highlight of RNTN is its capability of learning shifting directly from labeled training data without explicit knowledge of shifters and shifting rules. However, RNTN depends on manually labeled training data, i.e. sentences in which each node of the parse tree is labeled with polarity information. Such fine-grained manual annotation is currently only provided by the *Stanford Sentiment Treebank (SST)* (Socher et al., 2013). Resources like SST are not suitable for either training or testing a polarity classifier with respect to verbal shifters, since they do not contain each shifter with sufficient frequency. For example, SST contains instances for 16.9% of our verbal shifters, with less than half of these occurring more than once. We expect that RNTN, which has been trained on SST, may only be able to model shifting caused by frequently occurring negation words, but, unlike our own approach, will fail to account for shifting involving any but the most frequent verbal shifters.

Our own approach (**LEX**) is based on inferring the polarity of each VP from the polarity of the noun and whether the verb is a shifter. A VP with a shifter has a polarity moving to the opposite of the noun, a VP without shifter has the same polarity. We evaluate the shifter lexicons generated by our best graph-based classifier (**LEX$_{LP}$**) and best supervised classifier (**LEX$_{SVM}$**) from §5.2. Our

human annotated list of 980 shifters (§5.3) establishes an upper bound (**LEX$_{gold}$**).

Results in Table 9 show that all lexicons exceed the baselines. Even automatically induced shifter lexicons clearly outperform the prediction of existing sentiment analysis systems. Errors in LEX$_{gold}$ are mostly due to verbs that exhibit shifter behavior in some of their word senses, but not the one present in the phrase. In (17) *bring down* means *remove* and causes shifting, but in (18) its meaning of *inflict* does not cause shifting. The high scores produced by LEX$_{gold}$ also suggest that working on the lemma level instead of the sense level only means a moderate loss in performance.

(17) The revolution [[**brought down**]$_V$ the **tyrant**$_N^-$]$_{VP}^+$.
(18) She [[**brought down**]$_V$ a **curse**$_N^-$]$_{VP}^-$ on the village.

6   Related Work

Negation modeling is a central research issue in sentiment analysis, but only few works consider more than typical negation words. We refer the reader to the survey of Wiegand et al. (2010) for more information on negation modeling.

Approaches to learning negation from labeled corpora have been examined in the review domain (Ikeda et al., 2008; Kessler and Schütze, 2012; Socher et al., 2013; Yu et al., 2016), the biomedical domain (Huang and Lowe, 2007; Morante and Daelemans, 2009; Zou et al., 2013) and across domains (Fancellu et al., 2016). However, as outlined in §1, due to their small size the labeled datasets include few different verbal shifters. Moreover, these works mostly focus on scope detection rather than the identification of shifters.

The work most closely related to ours is Danescu-Niculescu-Mizil et al. (2009) who propose using NPIs for shifter extraction.[10] However, our work substantially extends that previous work. We show how the usage of NPIs can be further refined to improve the recognition of shifters (i.e. require the direct object to be a polar noun and subsequently apply PageRank). Moreover, we successfully combine this information with other features. Unlike Danescu-Niculescu-Mizil et al. (2009), we also carry out a recall-oriented evaluation and examine the impact of explicit knowledge of verbal shifters on contextual sentiment analysis.

---

[10]*Shifters* are referred to as *downward entailing operators*.

7 Conclusion

We took a major step toward producing a comprehensive lexicon of polarity shifters by bootstrapping a large list of verbal polarity shifters. Using a sample of 2000 manually annotated verbs extracted from WordNet, we built a supervised classifier to classify the remaining WordNet verbs. This reduced the number of verbs to be annotated manually by a large amount. We examined a variety of linguistic features and found that in addition to features derived from WordNet and FrameNet, the co-occurrence of the negative polarity item *any* with verbal shifters is particularly effective. We also showed that automatically learned knowledge of shifters improves the prediction of phrase-level sentiment.

Our approach should be largely transferable to other languages. This also applies to the features based on particular constructions such as the NPI *any*. The German NPI *jeglich*, Catalan *cap*, Japanese *dono mo* etc. can be expected to exhibit a very similar behavior (cf. Haspelmath (1997)).

Our goal is to build a complete lexicon of polarity shifters; to this end, future work will aim to add nouns and adjectives to our shifter lexicon.

Acknowledgements

References

E. Agirre and A. Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of EACL*.

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of COLING/ACL*.

E. Breck, Y. Choi, and C. Cardie. 2007. Identifying Expressions of Opinion in Context. In *Proceedings of IJCAI*.

L. Brinton. 1985. Verb Particles in English: Aspect or Aktionsart. *Studia Linguistica* 39:157–68.

Danqi Chen and Christopher D Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of EMNLP*.

Y. Choi and C. Cardie. 2008. Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. In *Proceedings of EMNLP*.

Y. Choi, L. Deng, and J. Wiebe. 2014. Lexical Acquisition for Opinion Inference: A Sense-Level Lexicon of Benefactive and Malefactive Events. In *Proceedings of WASSA*.

Y. Choi and J. Wiebe. 2014. +/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference. In *Proceedings of EMNLP*.

C. Danescu-Niculescu-Mizil, L. Lee, and R. Ducott. 2009. Without a 'doubt'? Unsupervised Discovery of Downward-Entailing Operators. In *Proceedings of HLT/NAACL*.

H. T. Dang. 2009. Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In *Proceedings of TAC*.

D. Das, N. Schneider, D. Chen, and N. A. Smith. 2010. Probabilistic Frame-Semantic Parsing. In *Proceedings of HLT/NAACL*.

D. Das and N. A. Smith. 2011. Semi-Supervised Frame-Semantic Parsing for Unknown Predicates. In *Proceedings of ACL*.

L. Deng, Y. Choi, and J. Wiebe. 2013. Benefactive/Malefactive Event and Writer Attitude Annotation. In *Proceedings of ACL*.

A. Esuli and F. Sebastiani. 2005. Determining the Semantic Orientation of Terms through Gloss Classification. In *Proceedings of CIKM*.

F. Fancellu, A. Lopez, and B. Webber. 2016. Neural Networks for Negation Scope Detection. In *Proceedings of ACL*.

L. Flekova and I. Gurevych. 2016. Supersense Embeddings: A Unified Model for Supersense Interpretation, Prediction, Utilization. In *Proceedings of ACL*.

A. Giannakidou. 2008. Negative and Positive Polarity Items: Licensing, Compositionality and Variation. In C. Maienborn, K. von Heusinger, and P. Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, Mouton de Gruyter, pages 1660–1712.

Y. Gyamfi, J. Wiebe, R. Mihalcea, and C. Akkaya. 2009. Integrating Knowledge for Subjectivity Sense Labeling. In *Proceedings of HLT/NAACL*.

S. Harabagiu, A. Hickl, and F. Lacatusu. 2006. Negation, Contrast and Contradiction in Text Processing. In *Proceedings of AAAI*.

K. S. Hasan and V. Ng. 2013. Frame Semantics for Stance Classification. In *Proceedings of CoNLL*.

M. Haspelmath. 1997. *Indefinite Pronouns*. Clarendon Press Oxford.

T. H. Haveliwala. 2002. Topic-Sensitive PageRank. In *Proceedings of WWW*.

Y. Huang and H. J. Lowe. 2007. A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. *Journal of the American Medical Infomatics Association* 14:304–311.

D. Ikeda, H. Takamura, L. Ratinov, and M. Okumura. 2008. Learning to Shift the Polarity of Words for Sentiment Classification. In *Proceedings of IJC-NLP*.

N. Jindal and B. Liu. 2008. Opinion Spam and Analysis. In *Proceedings of WSDM*.

T. Joachims. 1999. Making Large-Scale SVM Learning Practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, MIT Press, pages 169–184.

J. S. Kang, S. Feng, L. Akoglu, and Y. Choi. 2014. ConnotationWordNet: Learning Connotation over the Word+Sense Network. In *Proceedings of ACL*.

W. Kessler and H. Schütze. 2012. Classification of Inconsistent Sentiment Words using Syntactic Constructions. In *Proceedings of COLING*.

S. Kim, H. Chang, S. Lee, M. Yu, and J. Kang. 2015. Deep Semantic Frame-Based Deceptive Opinion Spam Analysis. In *Proceedings of CIKM*.

S. Kim and E. Hovy. 2006. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*.

S. Kiritchenko and S. M. Mohammad. 2016. The Effect of Negators, Modals, and Degree Adverbs on Sentiment Composition. In *Proceedings of WASSA*.

M. Krifka. 1991. Some Remarks on Polarity Items. In D. Zaefferer, editor, *Semantic Universals and Universal Semantics*, Foris Publications, pages 150–189.

J. R. Landis and G. G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1):159–174.

C. D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*.

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

M. Mitchell. 2013. Overview of the TAC2013 Knowledge Base Population Evaluation: English Sentiment Slot Filling. In *Proceedings of TAC*.

R. Morante. 2010. Descriptive Analysis of Negation Cues in Biomedical Texts. In *Proceedings of LREC*.

R. Morante and W. Daelemans. 2009. A Metalearning Approach to Processing the Scope of Negation. In *Proceedings of CoNLL*.

O. Sanchez-Graillet and M. Poesio. 2007. Negation of proteinprotein interactions: analysis and extraction. *Bioinformatics* 23(13):i424–i432.

R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. 2013. Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. In *Proceedings of EMNLP*.

V. Stoyanov and C. Cardie. 2011. Automatically Creating General-Purpose Opinion Summaries from Text. In *Proceedings of RANLP*.

G. Szarvas, V. Vincze, R. Farkas, and J. Csirik. 2008. The BioScope Corpus: Annotation for Negation, Uncertainty and their Scope in Biomedical Texts. In *Proceedings of BioNLP*.

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

P. P. Talukdar, J. Reisinger, M. Pasca, D. Ravichandran, R. Bhagat, and F. Pereira. 2008. Weakly-Supervised Acquisition of Labeled Class Instances using Graph Random Walks. In *Proceedings of EMNLP*.

M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo. 2010. A Survey on the Role of Negation in Sentiment Analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*.

M. Wiegand and J. Ruppenhofer. 2015. Opinion Holder and Target Extraction based on the Induction of Verbal Categories. In *Proceedings of CoNLL*.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of HLT/EMNLP*.

H. Yu, J. Hsu, M. Castellanos, and J. Han. 2016. Data-driven Contextual Valence Shifter Quantification for Multi-Theme Sentiment Analysis. In *Proceedings of CIKM*.

B. Zou, G. Zhou, and Q. Zhu. 2013. Tree Kernel-based Negation and Speculation Scope Detection with Structured Syntactic Parse Features. In *Proceedings of EMNLP*.