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

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 shear 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 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-Grailliet 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]¹⁻²].
(4) The new treatment has [alleviated,shifter her [pain]⁻¹²].

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. 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 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 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 expressions3 ranked by frequency (NEGATIVE).

---

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/ijcnnlp2017
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.\footnote{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.} 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 [Chavez] has fallen –effect\footnote{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.}.

(6) We don’t want the public getting the idea that we [abuse –effect our [prisoners] ]\footnote{\cite{565}.}.

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.\footnote{Polarity Clash (CLASH): Some of our previous examples (e.g. (3)) suggest that shifting is mainly caused by a polar verb (e.g. alleviate\textsuperscript{+}, deny\textsuperscript{+}) modifying a polar expression with the opposite polarity (e.g. [[\textit{lose}_\textit{shifter}]\textsuperscript{−}, [\textit{hope}_\textit{shifter}]\textsuperscript{+}], [[\textit{alleviate}_\textit{shifter}]\textsuperscript{+}, [\textit{pain}_\textit{shifter}]\textsuperscript{−}], or [[\textit{deny}_\textit{shifter}]\textsuperscript{−}, [\textit{scholarship}_\textit{shifter}]\textsuperscript{+}]). 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\textsubscript{norm}).} 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.

\footnote{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 [tore down\_shifter our great [dream] ]\textsuperscript{+}–.

(8) Please [lay aside\_shifter all your [worries] ]\textsuperscript{+}.

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 [not give us any [help\textsubscript{shifter}] ]\textsuperscript{+}–.

(10) They [denied\_shifter us any [help\textsubscript{shifter}] ]\textsuperscript{+}–.

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$_{\text{norm}}$). As a further constraint we demand that the direct object represents a polar expression (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, 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 on fish.
(12) Full keyboards on cellphones were first introduced in 1997.
(13) These buildings have been newly constructed.
(14) They specially prepared 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., 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.
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\text{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

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Freq</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>250</th>
</tr>
</thead>
<tbody>
<tr>
<td>FREQ</td>
<td>2000</td>
<td>90.0</td>
<td>82.0</td>
<td>78.0</td>
<td>78.0</td>
</tr>
<tr>
<td>POSITIVE</td>
<td>73</td>
<td>90.0</td>
<td>94.0</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>−EFFECT</td>
<td>95</td>
<td>90.0</td>
<td>92.0</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>ANTI</td>
<td>725</td>
<td>95.0</td>
<td>96.0</td>
<td>93.0</td>
<td>87.4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Acc</th>
<th>Prec</th>
<th>Rec</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline\text{majority}</td>
<td>84.8</td>
<td>42.4</td>
<td>50.0</td>
<td>45.9</td>
</tr>
<tr>
<td>kNN\text{noAntiShifter}</td>
<td>67.6</td>
<td>54.9</td>
<td>56.4</td>
<td>55.6</td>
</tr>
<tr>
<td>kNN</td>
<td>71.5</td>
<td>58.3</td>
<td>59.6</td>
<td>58.9</td>
</tr>
<tr>
<td>\text{LP}\text{noAntiShifter}</td>
<td>79.1</td>
<td>63.0</td>
<td>56.6</td>
<td>59.6</td>
</tr>
<tr>
<td>\text{LP}</td>
<td>80.7</td>
<td>68.6</td>
<td>56.7</td>
<td>62.0</td>
</tr>
<tr>
<td>\text{SVM}\text{task-spec. features} (§3)</td>
<td>79.9</td>
<td>63.5</td>
<td>69.7</td>
<td>67.5</td>
</tr>
<tr>
<td>\text{SVM}\text{generic features} (§4)</td>
<td>80.9</td>
<td>79.6</td>
<td>74.4</td>
<td>76.9</td>
</tr>
<tr>
<td>\text{SVM}\text{all features}</td>
<td>89.7</td>
<td>80.7</td>
<td>77.6</td>
<td>79.1</td>
</tr>
</tbody>
</table>

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

(i.e. ANY\text{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\text{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.
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 ANYnorm+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 ran 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 SVMlight (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.

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.**

<table>
<thead>
<tr>
<th>Conf. Rank</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-250</td>
<td>92.8</td>
</tr>
<tr>
<td>251-500</td>
<td>73.2</td>
</tr>
<tr>
<td>501-750</td>
<td>62.4</td>
</tr>
<tr>
<td>751-1043</td>
<td>33.2</td>
</tr>
</tbody>
</table>

\(^1\): final interval covers all remaining predicted shifters

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.
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 [lack, her usual passion,] [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 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; Kitichenko and Mohammad, 2016). The polar in-

<table>
<thead>
<tr>
<th>Information</th>
<th>Example</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence</td>
<td>Norah Jones’ smooth voice could soothe any savage beast.</td>
<td></td>
</tr>
<tr>
<td>Verb</td>
<td>soothe</td>
<td></td>
</tr>
<tr>
<td>Polar Noun</td>
<td>beast</td>
<td>negative</td>
</tr>
<tr>
<td>Verb Phrase</td>
<td>soothe any savage beast</td>
<td>positive</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Shifting Label</th>
<th>Noun Polarity ⇒ VP Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>shifted</td>
<td>+ ⇒ ∼ −</td>
</tr>
<tr>
<td></td>
<td>− ⇒ ∼ +</td>
</tr>
<tr>
<td></td>
<td>∼ ⇒ +</td>
</tr>
<tr>
<td>not shifted</td>
<td>+ ⇒ +</td>
</tr>
<tr>
<td></td>
<td>− ⇒ −</td>
</tr>
<tr>
<td></td>
<td>∼ ⇒ ~</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Acc</th>
<th>Prec</th>
<th>Rec</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Majority</td>
<td>79.9</td>
<td>39.9</td>
<td>50.0</td>
</tr>
<tr>
<td></td>
<td>RNTN</td>
<td>59.0</td>
<td>50.8</td>
<td>51.2</td>
</tr>
<tr>
<td>Proposed</td>
<td>LEXLP</td>
<td>84.3</td>
<td>77.3</td>
<td>67.4</td>
</tr>
<tr>
<td></td>
<td>LEXSYM</td>
<td>87.1</td>
<td>80.0</td>
<td>79.4</td>
</tr>
<tr>
<td></td>
<td>LEXgold</td>
<td>90.8</td>
<td>88.9</td>
<td>81.2</td>
</tr>
</tbody>
</table>

Table 9: Evaluation of polarity classification.
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]+]−/−.

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 (LEXLP) and best supervised classifier (LEXSVM) from §5.2. Our human annotated list of 980 shifters (§5.3) establishes an upper bound (LEXgold).

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 LEXgold 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 LEXgold 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 tyrantN]vp.
(18) She [[brought down]v a curseN]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. 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.

10Shifters 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

The authors would like to thank Stephanie Köser for annotating the verb lexicon presented in this paper. For proofreading the paper, the authors would also like to thank Anna Schmidt, Meaghan Fowlie, Anmearie Friedrich, Andrea Fischer, David M. Howcroft, Clayton Grenberg, Ines Rehbein and Katja Markert. The authors were partially supported by the German Research Foundation (DFG) under grants RU 1873/2-1 and WI 4204/2-1.

References


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


