Published in: Calzolari, Nicoletta/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Hasida, Koiti/Isahara, Hitoshi/ Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/ Odijk, Jan/Piperidis, Stelios/Tokunaga, Takenobu (eds.): Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 7-12 May 2018, Miyazaki, Japan. - Paris, France: European language resources association (ELRA), 2018. Pp. 608-612

Disambiguation of Verbal Shifters

Michael Wiegand*, Sylvette Loda*, Josef Ruppenhofer†
*Spoken Language Systems, Saarland University, †Institute for German Language
*D-66123, Saarbrücken, Germany, †D-68161, Mannheim, Germany
michael.wiegand@lsv.uni-saarland.de, sylvette.loda@lsv.uni-saarland.de
ruppenhofer@ids-mannheim.de

Abstract

Negation is an important contextual phenomenon that needs to be addressed in sentiment analysis. Next to common negation function words, such as *not* or *none*, there is also a considerably large class of negation content words, also referred to as *shifters*, such as the verbs *diminish*, *reduce* or *reverse*. However, many of these shifters are ambiguous. For instance, *spoil* as in *spoil your chance* reverses the polarity of the positive polar expression *chance* while in *spoil your loved ones*, no negation takes place. We present a supervised learning approach to disambiguating verbal shifters. Our approach takes into consideration various features, particularly generalization features.

Keywords: sentiment analysis, negation modeling, word-sense disambiguation

1. Introduction

Negation is one of the most central linguistic phenomena. Therefore, negation modeling is essential to various common tasks in natural language processing, 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). In the latter task, a negation typically reverses the polarity of polar expressions. For example, in (1), the negated positive polar expression *pass* conveys negative polarity.

(1) Peter did [**not** [pass]$^+$]$^-$ the exam.
(2) Peter [**failed** to [pass]$^+$]$^-$ the exam.

So far, most research in negation modeling for sentiment analysis has focused on negation (function) words, such as the particle *not* or the adverbs *no* or *never*. However, among other word classes, particularly the content words, such as verbs, nouns and adjectives, there also exist words expressing negation. For example, in (2) the verb *failed* has a similar function as the negation word *not* in (1). These *negation content words*, which are also called ***shifters***, are often excluded from discussion since there does not (yet) exist a commonly accepted resource with these expressions. Even though the frequency of a single negation function word is much higher than that of a shifter, the overall number of shifters is significantly larger than those of negation function words. Schulder et al. (2017) identified 980 verbal shifters while the popular negation word lexicon proposed by Wilson et al. (2005) only includes 15 negation function words. Moreover, since content words are more ambiguous than function words, we also envisage shifters to be more ambiguous than negation function words.

In this paper, we address the ambiguity of shifters. We select a set of 20 ambiguous verbal shifters and try to disambiguate them automatically. We focus on verbal shifters since it has been recently shown that a large amount of such verbs exist and they are important to polarity classification (Schulder et al., 2017). Examples for ambiguous verbal shifters are *clear*, *spoil*, *cloud* or *slump* which in (3), (5), (7) and (9) convey negation while in (4), (6), (8) and (10) they do not.

(3) The well-known actor was [**cleared** [of murder]$^-$]$^+$.
(4) That highly controversial bill *cleared* the House and caused three days of riots in the streets.
(5) Political blunders [**spoiled** [their chance of being re-elected]$^+$]$^-$.
(6) On Valentine's day, all people should *spoil* their loved ones.
(7) Suspicions of drug use have [**clouded** [her prospects of a promotion]$^+$]$^-$.
(8) The solution *clouds* if you shake it.
(9) [[My spirits]$^+$ **slumped**]$^-$ at the sight of him.
(10) After a busy day, the successful businessman *slumped* into his armchair and watched TV.

Our notion of negation focuses on the impact of polar expressions within the scope of a particular negation word or shifter. This may include cases that are no full negation in the proper semantic sense. For instance, in (5) a *chance* of being re-elected still exists. However, it has been significantly reduced. In terms of polarity, this is typically interpreted as a shift from positive to negative polarity (i.e. a *slight chance* is considered less positive than a *great chance*). It is not a proper negation in the sense that the there is no chance at all.

Moreover, the ambiguous verbs may also often carry an intrinsic polarity. Whether or not a particular mention of such verb conveys some form of negation, however, depends on the fact whether it shifts the polarity of its arguments. For example, in (5) the verb *spoil* acts as a shifter as the positive polarity denoted by its argument (i.e. *their chance of being re-elected*) is shifted. In (6), on the other hand, the verb conveys a positive polarity (as *spoil* here means treating someone with a lot of care and kindness). The polarity of its direct object is not reversed, i.e. the loved ones still remain to be loved. Therefore, in this sentence the verb does not function as a shifter despite carrying an intrinsic polarity.

In this paper, we follow a **supervised learning** approach. We frame the task as a binary classification problem. Either a mention of a verbal shifter actually conveys negation or it does not. The aim of this research is three-fold: First, we want to find out whether this disambiguation is learnable at all. Secondly, we want to determine which types of features are helpful for this task. Thirdly, we examine what type of training data is necessary: Is it necessary to have training data for every ambiguous verbal shifter in order to disambiguate them automatically or is it possible to generalize over different ambiguous verbs and therefore make automatic predictions for mentions of *unknown* ambiguous shifters?

All labeled data produced as part of this research effort are publicly available.[1]

2. Data & Annotation

We selected a set of 20 ambiguous verbal shifters (Table 1). We particularly focused on those types of verbs which display a high degree of ambiguity. This was established by annotating the senses associated with those verbs. All of our 20 verbs contain between 36.3% and 60.0% senses according to WordNet (Miller et al., 1990) in which the verb denotes some form of negation.

For each of those ambiguous verbs 100 sentences were randomly extracted from the North American News Text Corpus (LDC95T21) resulting in 2000 sentences. We only took sentences into account that had between 10 and 40 tokens. Short sentences, particularly headlines, and very long sentences are often incorrectly parsed. Since some of our features rely on the output of a syntactic parser and it would be beyond the scope of this work to improve parsing quality for very short and long sentences, we focus on those sentences that are more likely to be correctly parsed.

Each of the 2000 sentences was manually annotated. The annotator had to decide for each sentence whether the ambiguous verbal shifter which it contained conveys some form of negation or not. Instead of directly annotating the sentences from scratch, we first looked at the set of all possible word senses of each ambiguous verbal shifter (according to WordNet) and decided for each word sense whether it conveys negation or not. In the actual annotation of the sentences, the annotators also first determined the particular word sense (according to WordNet) it conveyed and then made the final decision on the basis of the assigned word sense. We found that by this procedure we could notably improve interannotation agreement. On a subset of 600 sentences, an interannotation agreement of $\kappa = 0.79$ was measured. This level of agreement can be considered substantial (Landis and Koch, 1977).

On the entire gold standard, 62% of the mentions of the ambiguous verbal shifters were annotated as shifters, while the remaining 38% were found not to convey any form of negation. Table 2 summarizes the most important statistics of our gold standard.

[1] https://github.com/miwieg/lrec2018

clear, corrupt, cost, crumple, depress, disdain, dissolve, drain, eject, hold down, jam, overturn, paralyse, sink, slump, soothe, spoil, trash, tumble, worsen

Table 1: The 20 ambiguous verbs.

Property	Freq
Number of unique ambiguous verbal shifters	20
Number of annotated sentences	2000
Number of sentences per verbal shifter	100
Average number of tokens in sentence	26
Proportion of mentions conveying negation	62%

Table 2: Some statistics of the gold standard.

3. Feature Engineering

3.1. Surface-based Features: Bag of Words

Since we can consider our task as some type of text classification task, we should examine simple surface-based features, such as bag of words. Despite both their genericity and simplicity, these features are known to be very predictive.

3.2. Word Generalization Features

Even though we expect bag of words to be predictive for this task, we also anticipate this type of feature to suffer from data sparsity. With 2000 sentences, our gold standard is not very large. The mere fact that we conduct a text classification on the sentence level further exacerbates this problem. (Sentences contain considerable fewer words than larger units of texts which are usually considered for text classification, such as paragraphs or documents.) Therefore, we examine different types of word generalization methods. What all these methods try to accomplish is that a classifier is able to classify a given context with words not observed in the training data. This can be achieved by harnessing the similarity of those unknown words and words observed in the training data.

Table 3 illustrates for the ambiguous verb *spoil* that the objects for the sense *negation* and *no negation* are of a specific semantic type. For the sense *negation*, the objects represent some form of activity, while for the sense *no negation*, they typically represent some human being (or at least some animate entity). These observed selectional preferences are an indication that some form of generalization of the context words might help for this task.

Brown Clustering. A popular data-driven word generalization method is Brown Clustering (Brown et al., 1992). This is an unsupervised clustering method in which words with the similar distributional contexts are automatically assigned the same clusters. Clusters therefore represent a set

Sense	Contexts	Type
negation	spoil a(n) {effort|idea|fun}	*activity*
no negation	spoil one's {partner|girlfriend|spouse}	*human*

Table 3: Illustration of word generalization.

[Negation Sense	No Negation Sense
[**clear** someone of murder^{-}]$^{+}$	sky *cleared*
[**spoil** someone's efforts^{+}]$^{-}$	*spoil* one's spouse
[peace^{+} **crumples**]$^{-}$	*crumple* a shirt
[**cost** someone their inner peace^{+}]$^{-}$	*cost* some amount of money

Table 4: Polar expressions. to indicate negation sense.

of words rather than individual ones.

In our experiments, we induce 1000 clusters from the North American News Text Corpus. This is a standard configuration proven to yield good results (Turian et al., 2010). We induce the clusters using the SRILM-toolkit (Stolcke, 2002).

Word Embeddings. A more recent alternative to Brown clustering is the usage of word embeddings. Word embeddings are (dense) vector representations of words that are automatically induced from corpora. They are devised as a more robust alternative to bag of words. Unlike a *one-hot* bag-of-words vector representation where different words (no matter how similar they are in meaning) are always orthogonal to each other, embeddings allow different words which are distributionally similar, such as *partner* and *spouse*, also to have similar vector representations. In our experiments, we induce word embeddings using Word2Vec (Mikolov et al., 2013). The embeddings are induced from the North American News Text Corpus. In order to avoid overfitting, we leave the tool in its default configuration.

WordNet Hypernyms and Supersenses. We use WordNet (Miller et al., 1990) as a resource-based method for word generalization. WordNet is the largest lexical ontology for the English language. It is organized in word senses called *synsets*.[2] By considering hypernyms of a synset representing some set of lemmas as a feature, we enable similar words, i.e. synonyms or near-synonyms, to have a feature in common. While hypernyms are a fairly fine-grained form of generalization, we also consider *supersenses*, a set of coarse-grained classes, which have previously been found to be effective for sentiment analysis (Flekova and Gurevych, 2016).

3.3. Polarity Information

We assume that many ambiguous shifters convey a negation if they co-occur with polar expressions. This is illustrated in Table 4. Therefore, we count the number of polar expressions in a sentence to be classified. We identify such expressions with the help of the *Subjectivity Lexicon* (Wilson et al., 2005).

3.4. Focused Features

Our task can be considered as a word-sense disambiguation (WSD) task. Therefore, we should also consider a feature set established in previous work on WSD. Table 5 lists those features mainly inspired by Akkaya et al. (2009).

[2]Since we are not aware of any robust open-domain word-sense disambiguation software, we always consider the union of all synsets associated with a particular lemma.

Feature
subcategorization frame of verbal shifter
hypernym(s) of dependents of verbal shifter
supersense(s) of dependents of verbal shifter
is a polar expression among dependents of verbal shifter?
is verbal shifter coordinated with another verbal shifter?
words representing dependents of verbal shifter

Table 5: Focused features

	unknown verbs		known verbs	
Feature	**Acc**	**F1**	**Acc**	**F1**
bag of words	63.8	59.6	70.7	67.9
embeddings	64.4	60.8	70.8	68.1
bag of words + embed.	64.3	60.7	70.9	68.2

Table 6: Bag of words vs. word embeddings.

They have in common that they all only consider a very local context of the mention of the verbal shifter (i.e. typically its dependents). Some of these features incorporate syntactic information. We use the Stanford Parser (Chen and Manning, 2014) to obtain that information.

4. Experiments

We evaluate our features in a 10-fold crossvalidation. The classifier, we consider is a Support Vector Machine. As a tool we use SVM^{light} (Joachims, 1999). We consider two different evaluation settings. On the one hand, we arrange the verbs in such a way that the test data contain contexts of verbs which have not been observed in the training data. On the other hand, we ensure that all test data contain contexts of verbs that have been observed in the training data. As a baseline, we also list the performance of a majority-class classifier.

For the bag-of-words features, we experimented with different window sizes but found that using all words in a sentence provides best performance. For the word embeddings, however, we had to consider a fixed window size because we need to establish that all vectors representing an instance have the same dimensionality. In order to achieve this, we took the word embeddings of the words in a fixed context window and simply concatenated them to a large vector. We established that for the setting using observed verbal shifters, the optimal window size is $n = 8$ and for the setting of unknown verbs, the size is $n = 6$.

Table 6 compares the performance of bag of words and word embeddings and their combination. The performance of the two representations is very similar and there is no significant benefit in combining them. Consequently, we will exclusively employ the bag-of-words feature in our remaining experiments.

Table 7 compares different kinds of feature sets. It shows that we can significantly outperform our bag-of-words baseline on the setting dealing with unknown verbs. The performance on this setting is also notably worse than on observed verbs. Still, even on the former setting, we manage to outperform the majority-class baseline. This is an important result as it indicates that in order to learn this

	unknown verbs		known verbs	
Feature	**Acc**	**F1**	**Acc**	**F1**
majority	61.7	38.1	61.7	38.1
bag of words	63.8*	59.6*	70.7*	67.9*
all features	**66.8***	**63.1***	**70.8**	**68.1**

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

Table 7: Comparison of different features.

	unknown verbs		known verbs	
Feature	**Acc**	**F1**	**Acc**	**F1**
all features	66.8	63.1	70.8	68.1
w/o bag of words	66.8	63.3	70.8	68.1
w/o brown clusters	66.7	63.2	70.2	67.5
w/o hypernyms	65.0*	60.7*	69.8	67.1
w/o supersenses	66.2	62.1	70.7	68.1
w/o polarity	66.7	63.2	70.9	68.2
w/o focused	66.1	63.2	70.9	68.1

*: significantly worse than *all features* using paired t-test ($p < 0.05$)

Table 8: Ablation experiment.

disambiguation we do not require sentences with all possible ambiguous verbal shifters. However, our results also suggest that for such a setting a more sophisticated set of features is necessary.

Table 8 presents the results of an ablation experiment in which we compare the performance of the full feature set with a feature set where one type of feature is removed. The table shows that most feature types do not convey unique information since if they are removed, classification performance usually only drops marginally. There is only one notable exception: on the setting using unknown verbs the omission of WordNet hypernyms causes a significant drop in performance.

Figure 1 shows a learning curve of the most important feature sets. Due to the limited space of this paper, we only display the setting using unknown verbs. Judging by the slope of the curve, we could expect further improvements of classification performance by adding more training data. The table also shows that when more than 40% of the training data are used, the entire feature set systematically outperforms the bag-of-words baseline.

5. Related Work

With regard to WSD, our work follows the strand of research that argues for a coarse-grained set of sense inventories (Palmer et al., 2004; Hovy et al., 2006; Navigli, 2006; Snow et al., 2007). Coarse-grained sense inventories have an obvious practical advantage. They are easier to discriminate than fine-grained sense inventories. This applies to both human and automatic categorization. They typically also require fewer training data.

The work most closely related to ours is Akkaya et al. (2009) in that a coarse-grained sense inventory is examined for sentiment analysis. That work proposes two senses for expressions potentially conveying

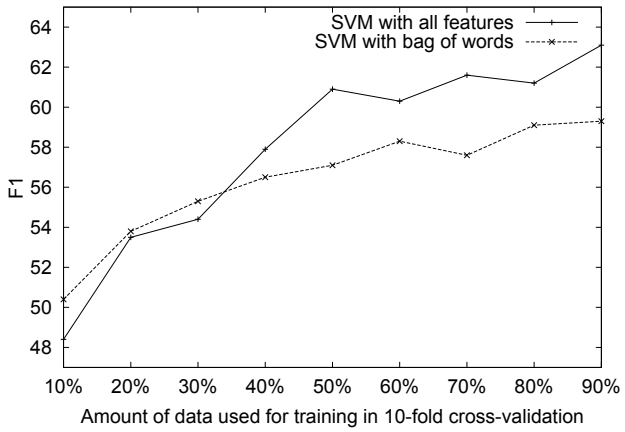

Figure 1: Learning curve on gold standard.

subjectivity, one sense in which the expression indeed conveys subjectivity and another in which it does not. We also consider two sense categories: a verb may convey negation or it may not.

To the best of our knowledge, this paper is the first work to address the ambiguity of shifters as a major research task. The most comprehensive study on negation with respect to polarity (Wilson et al., 2005) already identified the problem of negation words being ambiguous. Since the negation words that are considered in that work are predominantly negation function words, the problem is not considered that severe. In fact, those few cases of ambiguity that are identified concern the negation particle *not* and can be reliably disambiguated by a handful of simple phrasal patterns, that is, *not only* in (11), *not just* in (12) and *if not* in (13). Note that the polar expressions in the scope of the mentions of *not* in these three examples (i.e. *problem* in (11), *skill* in (12) and *worse* in (13)) all preserve their polarity in those contexts, that is, they are not negated by *not*. For our verbal shifters, such simple phrasal rules could not be identified on our dataset.

(11) This is *not* only a problem$^-$ that concerns this nation but the entire world population.

(12) Camouflage is *not* just a critical skill$^+$ in combat, it is the recipe for victory.

(13) The situation in Afghanistan is as bad, if *not* worse$^-$ than, under the Taliban.

6. Conclusion

We presented an approach to automatically disambiguate verbal shifters. The task was framed as a supervised learning approach. We found that, in principle, one can learn to distinguish these senses and that even classification of unknown shifters is possible. Here, particularly word generalization features are important. In general, very simple surface features, such as bag of words are already effective. Our learning curve suggests, however, that in order to produce a classifier with reasonable performance more labeled training data are required.

Acknowledgements

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

7. Bibliographical References

Akkaya, C., Wiebe, J., and Mihalcea, R. (2009). Subjectivity Word Sense Disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 190–199, Singapore.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

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.

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.

Harabagiu, S., Hickl, A., and Lacatusu, F. (2006). Negation, Contrast and Contradiction in Text Processing. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 755–762, Boston, MA, USA.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 57–60, New York, NY, USA.

Joachims, T. (1999). Making Large-Scale SVM Learning Practical. In B. Schölkopf, et al., editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. MIT Press.

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

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

Miller, G., 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.

Navigli, R. (2006). Meaningful Clustering of Senses Help Boost Word Sense Disambiguation. In *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, pages 105–112, Sydney. Australia.

Palmer, M., Babko-Malaya, O., and Dang, H. T. (2004). Different Sense Granularities for Different Applications. In *Proceeddings of the HLT-NAACL-Workshop on Scalable Natural Language Understanding*, pages 49–56, Boston, MA, USA.

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

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)*, Taipei, Taiwan.

Snow, R., Prakash, S., Jurafsky, D., and Ng, A. Y. (2007). Learning to Merge Word Senses. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1005–1014, Prague, Czech Republic.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904, Denver, CO, USA.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word Representations: A Simple and General Method for Semi-supervised Learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 384–394, Uppsala, Sweden.

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*, pages 60–68, Uppsala, Sweden.

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