Finding Domain Specific Polar Words for Sentiment Classification

Mehrbod Sharifi
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
msharifi@cs.cmu.edu

William Cohen
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
wcohen@cs.cmu.edu

Abstract

This paper presents a method of using conditional random fields (CRF) for extracting polar words and determining the overall sentiment of text. We frame sentiment classification as a feature selection problem and conduct three sets of experiments by using: prior polarity lexicons, bag-of-words classifiers and CRF sequence models. The results show the potential of utilizing CRFs in discovering high quality context-dependent polar features.

1 Introduction

Sentiment classification is the task of determining the sentiment of text (user reviews, blogs, broadcast news, etc.). The scope can be the overall sentiment of the document (Pang and Lee, 2002; 2004; 2005), or it can be focused on finding specific sentiment toward an object or entity (Hu and Liu, 2004; Popescu and Etzioni 2005). The output can be binary (positive or negative) or expressed in finer granularity (e.g., 5-star or 10-star rating).

A much related topic is the task of finding polar words, i.e., words that express an opinion. Hatzivassiloglou and McKeown (1997) showed how polarity of adjectives can be inferred from conjunctions and disjunctions in a corpus. Turney (2002) used web search to estimate point-wise mutual information (PMI) between a seed set of known polarity and any phrase to determine semantic orientation. Hu and Liu (2004) started from a small seed set and then extended it using synonym and antonym relations WordNet. From previous research, several "prior polarity lexicons" have been made available which we will utilize in our experiments. These resources indicate the polarity of words regardless of context, however it should be noted that word polarity can change in context (Wilson et al. 2005).

For sentiment classification, we need to combine word polarities and make a decision about the sentence (or document) polarity. Various methods have been explored: Hu and Liu (2004) take the majority of the positive and negative words (considering any negation in proximity). Kim & Hovy (2004) experimented with product of polarity signs and the arithmetic and the harmonic means of sentiment strength score. Popescu and Etzioni (2005) used relaxation labeling where they assigned polarity in three stages (word, phrase and sentence) and propagated information from each stage in optimization process.

While polar words are important clues to overall polarity, the context can heavily affect the accuracy. Pang and Lee (2002) bring up this problem with their bag-of-word classifiers: one class of errors ("thwarted expectation") is when the reviewer starts by many negative sentences but at the end negates the whole review with a short sentence of "but I liked it!". Indeed by error analysis of SVM, we see the majority of classifier errors exhibits some complexity resulted from sequence. We will not attempt to enumerate too many of these problem but will try to highlight a few. In "a great script brought down by lousy direction" coexistence of "great" and "lousy" can cause errors, but the sequence and verb "brought down" can be clues for determining correct polarity. Sometimes these clues are more complex but still sequential: "a respectable but uninspired thriller that's in-

telligent and considered in its details, but ultimately weak in its impact". One example of common patterns is the change of sentence sentiment around the word "but" (e.g., above and "interesting, but not compelling."). The SVM error rate in sentences containing "but" is much larger[1]. Some attempts have been made to capture negation in context for a bag-of-words approach, but the improvement reported has been minimal (Pang and Lee, 2002).

2 Approach

Since polarity is closely related to context, we propose using sequence models for incorporating it into classification. In particular, we use conditional random fields (CRF) for extracting the polar word. CRFs are a form of undirected graphical models used in modeling the sequential data (Lafferty et al. 2001). They have been successfully used in many tasks (such as Part-of-speech tagging or named entity recognition) and shown to generally outperform other models such as HMMs (CRF makes fewer independence assumptions).

To create and test a CRF polarity extractor, we need to have data in which the individual words are tagged with polarity. Available polarity annotations are often for phrases and other purposes and they unfortunately were not suitable for our experiments. We will experiment with using the prior polarity and classifiers to create the data for the CRF. In this process, we also evaluated the utility of these methods in the classification.

Classification given a set of feature weights will be done using this formula:

$$S(x) = sign\left(\sum_{i=1}^{n} f_i(x)\right) \qquad (1)$$

Where S is +1 for positive and -1 for negative reviews. $f_i(x)$ is the feature score extracted from the vector[2] x and total of n features are used.

2.1 Prior Polarity Lexicons

In this method, features will be selected if they are in the lexicon. Three lexicons have been used: General inquirer (2000) consist of 4200 entries, Subjectivity Clues (Wilson and Wiebe, 2005) with 8220 entries and SentiWordNet (Esuli and Sebastiani, 2006) with sentiment score for each WordNet entry (more 80K of unique words). For first two, -1 or 1 is used as the feature score. Any repetition within these sets is collapsed by averaging the scores.

2.2 Feature Selection with Classifier

Machine learning classifiers are precisely designed to perform this feature selection and weighing. This corresponds to finding a hyperplane which separates the two classes of reviews by optimizing a certain loss function (e.g., hinge loss in case of SVM). Most classifiers can be expressed as (1) with some simple transformation. For example, if we used a Naïve Bayes classifier, then the feature score can be maximum likelihood of (log) conditional probability distribution of words given the sentiment of review. However, Naïve Bayes is making a feature independence assumption, which is what we set out to remove[3]. Boosting (Shapire, 2003) has been shown as a good feature extraction method and it is not making any independence assumptions. We set $f_i = a_i h_i(x)$ and n as number boosting rounds ($h_i(x)$ are decision stump that splits training set based on one feature picked to minimize weighted training error).

3 Experiments

3.1 Dataset

Two datasets are used for our experiments[4]: polarity dataset v2.0 (Pang and Lee, 2004) which we will refer to as long reviews (1000 positive and 1000 negative reviews, avg. 780 words) and sentence polarity dataset v1.0 (Pang and Lee, 2005) henceforth short reviews (5331 positive and 5331 negative reviews, avg. 21 words).

3.2 Results

[1] The word "but" appeared in 16% of the short review test set (see experiments) and SVM error was 35% vs. 19% for whole set.

[2] This is a general definition and vector is not a bag-of-word representation of the review. This allows for a more complex hyperplanes to be generated.

[3] Use of higher n-gram is shown not to be effective and sometime harmful (Pang and Lee, 2002)

[4] http://www.cs.cornell.edu/People/pabo/movie-review-data/

	Long	Ties	Short	Ties
Subjectivity Clues	61.0%	<1%	70.9%	18.5%
General inquirer	56.6%	<1%	70.4%	23.2%
SentiWordNet	56.2%	<1%	59.9%	2.4%

Table 1: Accuracy Using Prior Polarity Lexicons

The result of experiments using three prior polarity lexicons is shown in Table 1. Depending on how hard we penalize ties, we see some improvements over the 50% baseline (random or all-in-one-class), but this is significantly lower than SVM[5]. Note that this result is with minimal processing for applying the prior polarity scores as feature weights. It is possible to achieve better accuracy through more analysis, e.g., finding the correct word sense when using SentiWordNet.

For feature extraction experiments, boosting was trained on short reviews[6]. A total of 3462 features were extracted and weights are set as explained in 2.2. To find a smaller set of features, we can consider the features with absolute score values greater than a threshold. Figures 1 shows the classification accuracy and Figure 3 shows the number of feature for various threshold values. Ties shown on the figure are generally 0-0 (when none of the review words were recognized as polar in the feature set).

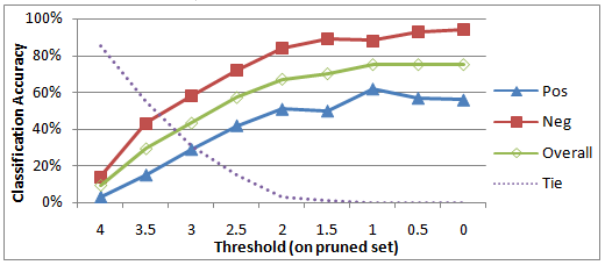

Figure 1: Accuracy of Pruned Boosting Features on Long Review (Baseline 50%)

Full set: *Pos* (outlandish, moodiness, liberating, combine, shrek), *Neg* (bore, dogs, blank, stunt, disappointment)
Pruned: *Pos* (moodiness, combine, fulfill, priceless, mesmerizing), *Neg* (blank, stunt, disappointment, brawny, whiny)

Figure 2: Top 5 Features from boosting

[5] Boolean features vector, discarding features with frequency lower than 4, no stemming or stop word removal
[6] 1000 rounds reached 76% accuracy. More training rounds or use of n-grams up to n=6 didn't have significant improvements or lowered the accuracy.

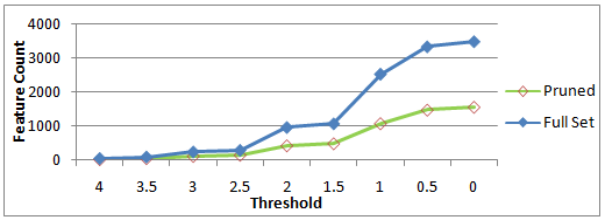

Figure 3: Feature counts for each threshold level

Token Counts		CRF			
		POS	**NEG**	**O**	**Total**
Pruned	**POS**	3275	1	205	3481
	NEG	5	3787	251	4043
	O	**38**	**29**	51599	51666
	Total	3318	3817	52055	59190

Table 2: Contingency table for CRF Result

Some of the highly scored features are shown in figure 2, but they are not polar words (e.g., "dogs" or "shrek" are part of movie names that generally received positive and negative reviews). We consider "pruning" these features by intersecting the full set with a prior lexicon (SentiWordNet, because it contained more words). Pruning has minimal effect on the accuracy of the classifier but the number of feature are considerably reduce (figure 3). This is a critical step and using the full set of feature will produce worse results in training CRF. We believe this is because the words that are not polar do not share the same contextual clues as the polar words and will confuse the classifier.

Table 2 summarizes the result of training CRF to extract additional features and it shows a high precision and recall for polarity. The training sets has been generated by tagging 2/3 of short reviews with pruned boosting features with 1 threshold (around 1000 polar words). Only reviews with at least one polar word are included. To ensure that CRF is generalizing (not overfitting to training set) we observed that extractor has discovered 18% more polar words than the training set including 8% completely new word (i.e., correct – by introspection – polar words not included in the original polarity set used for tagging).

CRF training is sensitive to noise and therefore using the full set or lower weight threshold resulted in less accurate extractor. This currently limits our ability to extract many additional polar words and hence the classification improvement was very small (1%-2%) when we added the features extracted by CRF to our full set. The other problem

to be considered is the weight in which the newly discovered polar word should be added to the original feature set.

4 Related Work

Use of sequence modeling for context has mostly been in information extraction and not for polarity detection. Some methods used context differently: Wilson, et al. (2005) did phrase level sentiment classification by extracting features from the context of words with prior polarity to find their contextual polarity. Rilof and Wiebe (2003) used bootstrapping to learn extraction patterns. Popescu and Etzioni (2005) followed a similar approach but also incorporated PMI measure. These patterns are fixed syntactic patterns and are quite different than sequence probability distributions learned by CRF. The most relevant to our work was Mao and Lebanon (2007): they created special (isotonic) CRF to model the flow of sentiment in a sentence. They also bring up the issue of using the CRF on ordinal label sequences. We do not use CRF for classifying the sentiment of the review but to extract polar words and then use them to improve classification.

5 Conclusion and Future Work

We showed results of experiments with different feature selection methods: three prior polarity lexicons, boosting as a bag of words classifier and conditional random fields as a sequence model. This opens up many aspects of this problem to be explored. We can improve how we applied prior polarity. We also need to understand further what exactly should be modeled as patterns for polarity. This may allow us to train better CRFs or encode the clues in a different form than CRF (which may allow more linguistic information to be embedded as well).

We also consider running experiment to understand how human selects the polar word and their weight from the contextual clues.

References

Esuli, Andrea and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *In Proceedings of LREC-06, 5th Conference on Language Resources and Evaluation*, Genova, IT.

The General Inquirer. 2000. http://www.wjh.harvard.edu/~inquirer

Hatzivassiloglou, Vasileios and Kathleen McKeown. 1997. Predicting the Semantic Orientation of Adjectives. *Proceedings of 35th Annual Meeting of the Assoc. for Computational Linguistics (ACL-97).* 174-181

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews". *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004)*, Seattle, Washington, USA.

Kim, Soo-Min and Eduard Hovy. 2004. Determining the Sentiment of Opinions. *Proceedings of COLING-04*. pp. 1367-1373. Geneva, Switzerland.

Lafferty, John, Fernando Pereira, and Andrew McCallum. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In International Conference on Machine Learning (ICML2001).*

Mao, Yi and Guy Lebanon. 2007. Isotonic Conditional Random Fields and Local Sentiment Flow. *Advances in Neural Information Processing Systems 19.*

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. *In Proceedings of EMNLP*, pages 79–86.

Pang, Bo and Lillian Lee. 2004. A sentiment education: Sentiment analysis using subjectivity summarization based on minimum cuts. *In Proceedings of the ACL*, pages 271–278.

Pang, Bo and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *In Proceedings of the ACL*. pages 115--124.

Popescu, Ana-Maria and Oren Etzioni. 2005. Extracting Product Features and Opinions from Reviews. *In Proceedings of HLT-EMNLP 2005.*

Riloff, Ellen and Janyce Wiebe. 2003. Learning Extraction Patterns for Subjective Expressions. *In Proceedings of EMNLP*. Pages 105-112.

Schapire, Robert E. 2003. The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, B. Yu, editors, *Nonlinear Estimation and Classification. Springer, 2003.*

Turney, Peter. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classication of reviews. *In Proceedings of the ACL*, pages 417–424.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *In Proceedings of HLT-EMNLP 2005.*