Detecting Domain Dedicated Polar Words

Raksha Sharma, Pushpak Bhattacharyya
Dept. of Computer Science and Engineering
IIT Bombay, Mumbai, India
{raksha,pb}@cse.iitb.ac.in

Abstract

There are many examples in which a word changes its polarity from domain to domain. For example, *unpredictable* is positive in the movie domain, but negative in the product domain. Such words cannot be entered in a "universal sentiment lexicon" which is supposed to be a repository of words with polarity invariant across domains. Rather, we need to maintain separate domain specific sentiment lexicons. The main contribution of this paper is to present an effective method of generating a *domain specific sentiment lexicon*. For a word whose domain specific polarity needs to be determined, the approach uses the Chi-Square test to detect if the difference is *significant* between the counts of the word in positive and negative polarity documents. We extract 274 words that are polar in the movie domain, but are not present in the universal sentiment lexicon. Our overall accuracy is around 60% in detecting movie domain specific polar words.

1 Introduction

Sentiment analysis (SA) has attracted a great deal of attention in recent times (Hatzivassiloglou and McKeown, 1997; Wiebe, 2000; Pang et al., 2002; Turney, 2002; Yu and Hatzivassiloglou, 2003; Hu and Liu, 2004; Esuli and Sebastiani, 2005; Breck et al., 2007). The SA task is to predict the sentiment orientation of a text (document/para/sentence) by analyzing the polarity of words present in the text. A lexicon of sentiment bearing words is of great help in such tasks.

Sentiment lexicons are of two types: universal and domain specific. Words like 'good' and 'bad' have uniform polarity across all domains, and so are members of universal sentiment lexicon. A word like 'unpredictable', on the other hand, is positive in the movie domain ('unpredictable plot'), but negative in the car domain ('unpredictable steering'). Such a word should be entered as positive in the sentiment lexicon of the movie domain and as negative in the sentiment lexicon of the car domain.

There are many "universal sentiment lexicons" like SentiWordNet[1], subjectivity lexicon[2] by Wiebe, list of positive and negative opinion words[3] by Liu. These lexica contain only those polar words which have the same polarity in all domains. In this paper, we use the universal sentiment lexicon published by Wiebe.

Using resources like Wikipedia and SentiWordNet to determine polarity of a domain specific word may lead to wrong sentiment detection. The motivation for our work comes from addressing this problem. We would like to create domain specific sentiment lexicons.

Our technique for detecting domain specific polar words is inspired by the work done by Cheng and Zhulyn (2012). They used the Pearson's Chi-Square test to find the top 200 words most indicative of positive sentiment and the top 200 words most indicative of negative sentiment from the corpus itself. They used these words as the lexicon for the hitting 2-gram language model. They observed that the hitting 2-gram model achieves far greater accuracy than other language models. In their work, they used the categorical Chi-Square test to determine the score of a word with positive sense and negative sense. Their Chi-Square test gives weightage also to those documents in which the word is absent while calculating the score. However, their idea of selecting hitting words, considers multiple occurrences of a word in a single doc-

[1] http://sentiwordnet.isti.cnr.it/
[2] http://mpqa.cs.pitt.edu/
[3] http://www.cs.uic.edu/liub/FBS/sentiment-analysis.html

International Joint Conference on Natural Language Processing, pages 661–666,
Nagoya, Japan, 14-18 October 2013.

ument as one. This leads to loss of information that can help in deciding the correct polarity of a word from the corpus. We use the *goodness of fit Chi Square* test, that takes into account the total occurrences of a word in the corpus to assign the score. This test allows us to compare a collection of categorical data with some theoretical expected distribution[4].

Our proposed method identifies sentiment words from the corpus. Wiebi (2000) observes that the probability of a sentence being subjective given that there is at least one adjective in the sentence is 55.8%. So we presently focus on adjectives. The key idea is that if a word can have both positive or negative polarity, then it should be uniformly distributed between positive and negative files. For this purpose, we take an equal number of positive and negative reviews from the same domain. So, the expected count of the word in positive and negative reviews is half of the total count in the corpus. This is the null hypothesis.

If the word satisfies the Chi-Square test, it indicates that there is a significant difference between the expected and observed count of the word. Hence, the null hypothesis should be rejected and it should be considered that this deviation from expected value is not by chance, but because of the domain specific polarity of the word, which makes the word more frequent in one of positive or negative reviews.

The road map for the rest of the paper is as follows: Section 2 describes the previous work done in the direction of sentiment lexicon. Section 3 elaborates on the generation of domain specific polar words through the Chi-Square test. Section 4 gives the experimental set up. In section 5, we present results along with discussions. We conclude the paper with points for future work in section 6.

2 Related Work

Extensive work has been done in the area of universal sentiment lexicon using corpora based approaches. Wiebe (2000) focused on the problem of identifying subjective adjectives with the help of the corpus. They proposed an approach to find subjective adjectives using the results of a method for clustering words according to their distributional similarity, seeded by a small number of simple adjectives. These adjectives were extracted from a manually annotated corpus. The basic idea is that subjective words are similar in distribution as they share pragmatic usages. However, the approach is unable to predict sentiment orientations of the found subjective adjectives.

Some evidence exists in the area of domain specific sentiment lexicon. The work of Kanayama and Nasukawa [2006], demonstrates the extraction of domain specific sentiment words in Japanese text. They exploited clause level context coherency to find candidate words for domain specific sentiment lexicon from sentences that appear successively with sentences containing a word from the seed set. The seed set is the set of strong universally polar words. The intuition is that sentences appearing in contexts tend to have the same polarities, so if one of them contains sentiment words, the other successive sentences are expected to contain sentiment words too, with the same polarity. Then, they use a statistical estimation based method to determine whether the candidates are appropriate sentiment words. However, the idea of using a seed set to extract purely domain dependent words may lead to wrong polarity.

Qiu et al. (2009) exploited the relationship between sentiment words and product features that the sentiment words modify in a domain dependent corpus. They used sentiment words and product features to extract new sentiment words. The extraction rules are designed, based on relations described in dependency trees. Their method also begins with a seed set. They proposed that a feature should receive the same polarity in a review and the words extracted by this feature will receive polarity of feature. However, the reviewer may associate polarity with a feature of time. If time changes, his views for a feature may change in the same review. To understand this fact consider the following example.

> *"When I purchased this camera, the battery was good, but now it is disastrous".*

Qui et al. (2009) considered 'camera', 'DVD player', and 'MP3 player' as one domain. However, *grainy* and *blurred* are negative in the camera domain, but neutral for 'DVD' and 'MP3 player'. Our work is independent of features and the seed list. It only needs a sufficient equal number of positive and negative review files written by a reliable source.

[4] http://math.hws.edu/javamath/ryan/ChiSquare.html

3 The Proposed Method

In this paper, we focus on finding sentiment words for the movie domain with their polarity as positive or negative. Finding movie domain specific polar words is an appealing task for several reasons. First, providing polarity information about movie reviews is a useful service. Its proof is the popularity of several film review websites [5]. Second, movie reviews are harder to classify than reviews of other products (Turney, 2002) and so is the classification of sentiment words. Our data contains 1000 positive and 1000 negative reviews, all written before 2004[6]. Movie reviews are accompanied by plot descriptions and plot is not a part of the reviewer's opinion of the movie. So presence of polar words in plot description can mislead the Chi-Square test. To solve this problem, we clean the corpus by removing the plot description from reviews, before giving it as an input to the Chi-Square test. Cleaning of the corpus is done automatically by finding patterns for plot description in movie reviews. In this paper, we perform the Chi-Square test with adjectives extracted from both cleaned and non cleaned corpus. The orientation of polarity of the output sentiment words are predicted simultaneously.

3.1 Sentiment Word Extraction and Polarity Assignment

The key idea is that if a word does not belong to a particular class, then it should be uniformly distributed among all classes. So, before starting the test, we consider a null hypothesis. A null hypothesis states that if a given word is neutral, its chance to occur in positive and negative documents is equal. The value of a null hypothesis is equal to the arithmetic mean of the word count in positive and negative documents. This can also be considered as the expected count of words in both the classes of documents. We apply the Chi-Square test on the expected count and the actual observed count of the word. Deciding the polarity of words, that are used very rarely in corpus, is not worth considering. Since, if a word is polar, then it will occur frequently in polar documents. So we give only those words as input to the Chi-Square test, whose mean value is greater than 6. If the Chi-Square test results in a value, which is greater than the threshold value, then there is a significant difference between the expected and the observed count of the word. At this moment we reject the null hypothesis and consider the possibility that there is some other factor causing the observed count to differ from the expected count of words. This factor is nothing but the polarity of adjectives, which makes it appear in a particular type of documents, frequently. If the word has positive sentiment, then it will occur more frequently in positive documents. Consider the following example.

> *mesmerizing, unpredictable, thrilling, non-stop*

Negatively polar words occur more frequently in negative documents. Consider the following example.

> *juvenile, predictable, underwritten, murky*

The extraction approach is best described in Algorithm 1.

The Bidirectional Standford POS tagger[7] is used to tag words from the corpus with parts of speech. Experiments are performed with different thresholds for the Chi-Square value of the adjective.

3.2 Cleaning of Corpus

In the movie domain, reviewers feel free to describe the plot of the movie as part of the review for a better understanding of it. So, in the movie domain, the *cleaning of the corpus* is mandatory because the polar words which are present in the plot part may mislead the classifier. However, cleaning of the corpus is not required in other domains, for example, *Camera* and *Cell Phones*. We find patterns that represent plot description in the corpus.

- Some reviewers have explicitly divided reviews into two parts - one for review and another for the movie plot - under different titles. It is shown in table 1.
- Some reviewers have specified that the *review contains spoilers.*
- We are performing experiments with the *English* movie review corpus, so movie names

[5] www.rottentomatoes.com, www.imdb.com

[6] Available at www.cs.cornell.edu/people/pabo/movie-review-data/ (review corpus version 2.0)

[7] http://nlp.stanford.edu/software/tagger.shtml

Input: Domain Specific Corpus Tagged with POS
Output: Sentiment Lexicon with Polarity
foreach $WORD$ *in the corpus* **do**
  if *POS of* $WORD$ *is* JJ *or* JJS **then**
    T:= get total count($WORD$)
    P:= get count in positive documents($WORD$)
    N:= get count in negative documents($WORD$)
    Expected_Count := $T/2$;
    if $Expected\_Count > 6$ **then**
      $Chi^2(WORD) :=$ $(((P - Expected\_Count)^2 +$ $(N - Expected\_Count)^2)/$ $Expected\_Count)$
      if $Chi^2 > Threshold$ **then**
        if $(P - N) > 0$ **then**
          Polarity := +1;
          Add_To_Sentiment_Lexicon ($WORD$,Polarity);
        else
          Polarity := -1;
          Add_To_Sentiment_Lexicon ($WORD$,Polarity);
      else
        Continue for next $WORD$;
    else
      Continue for next $WORD$;
  else
    Continue for next $WORD$;
end

Algorithm 1: Extraction of sentiment lexicon with the polarity

may overlap with adjectives, for example, *unhappy birthday, 13th warrior*. In a few places in reviews, movie names are given inside *double quotes*.

We find such files that match the pattern described above automatically and delete the found pattern.

4 Experimental Setup and Discussion

We use customer review collection as input data. The collection contains 1000 positive reviews and 1000 negative reviews. Experiments are done with cleaned and non cleaned corpora. We perform experiments with three threshold values 1.07, 2.45, 3.84. The threshold value specifies the minimum probability[8] to accept a null hypothesis. For example, a threshold value of 1.07 indicates that there must be more than a 30% probability, to accept a null hypothesis. If the Chi-Square value of a word is greater than 1.07, we can conclude from the Pearson Chi-Square probability table that there is less than 30% probability, to accept a null hypothesis. Hence, reject the null hypothesis and consider word as candidate for sentiment lexicon.

Plot Part	Review Part
Plot	Critique
Synopsis	Comment
Synopsis	Reviews
Ingredient	Opinion

Table 1: Parts of a Review

1.07 also classifies boundary words, whose sentiment is not very clear from the corpus. Boundary words are those words that have almost equal occurrence in positive and negative documents, since they occur less frequently in the whole corpus. So such words fail to qualify the Chi-Square test with higher threshold values, but are actually polar. With threshold values, 2.45, 3.84 we get an increment in precision at the cost of leaving some boundary words unclassified.

In one of the experiments, we were able to retain words with poor Chi-Square value and higher threshold, that is-3.84, with the help of universal sentiment lexicon. Universal sentiment lexicon contains words which are strongly polar independent of the domain(Wilson et al., 2005). If a word has been rejected by the Chi-Square test with a threshold of 3.84, and it belongs to universal sentiment lexicon, then the correct polarity of the word can be derived from universal sentiment lexicon. Consider the following examples.

> ***Distracting*** *gets a Chi-Square value 2.0 but certainly negative in all domains.*

> ***Monotonous*** *gets a Chi-Square value 2.25 but certainly negative in all domains.*

5 Results and Discussion

Since there is no gold standard sentiment lexicon for the movie domain, the quality of output obtained through the Chi-Square test is confirmed by the inter annotators agreement. We ex-

[8] http://faculty.southwest.tn.edu/jiwilliams/probab2.gif

tracted 11,828 adjectives from corpus as candidates for lexicon. Among them 932 adjectives fulfill the Chi-Square test on non-cleaned corpus with threshold 1.07. **476 adjectives** are marked as true positives by **inter annotators agreements**. Table 2 shows the precision obtained with non cleaned corpus. With a threshold of 1.07, we get a precision of 51%. This result is affected by the words that occur in the plot description. Words which are part of the plot description mislead the classifier, causing low precision. Table 3 shows an improvement in precision with the cleaning of the corpus. The Chi-Square test with a threshold of 3.84, and universal sentiment lexicon gives a very high precision, that is 69.1%.

With a small threshold of 1.07, we are able to fetch almost all the words from the corpus that can be candidates for sentiment lexicon in the movie domain. With this intuition, we use true positives (476) and false positives (456) extracted by the Chi-Square test with threshold of 1.07 on the non cleaned corpus as a gold standard data to calculate recall and accuracy for experiments whose results are shown in table 3.

Data Set	Threshold	Precision
Non-Cleaned Corpus	1.07	51.07%
Non-Cleaned Corpus + Universal Sentiment Lexicon	3.84	69.1%

Table 2: Precision of the Chi-Square test with a non-cleaned Corpus

Table 3 shows results of precision, recall with increasing threshold values for the Chi-Square test.

Threshold	Precision	Recall
1.07	54%	100%
2.45	59%	82%
3.84	61%	65%

Table 3: Precision of the Chi-Square test with a cleaned Corpus

Table 3 shows that, as the value of the threshold increases, the precision increases. However, recall decreases. Figure 1 shows the accuracy obtained with different Chi-Square threshold values. The words which have a good Chi-Square score are strong candidates for sentiment lexicon. But the words which are actually polar in the movie domain, but have been used very occasionally by reviewer, get rejected with an increase in the value of the threshold.

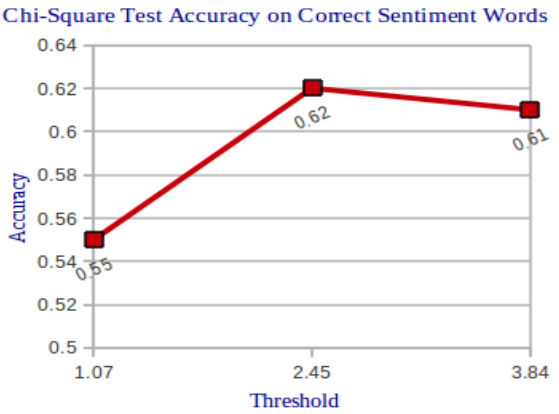

Figure 1: Chi-Square Test Accuracy with Different Thresholds

From figure 1, we can observe that accuracy is highest with a threshold of 2.45. When we move towards a higher threshold values, accuracy starts decreasing because of the higher fall in recall.

6 Conclusion

In this paper, we proposed a scheme to detect domain-dedicated sentiment words from the corpus. Our algorithm identifies polar words through an innovative application of Chi-Square test on the difference in the counts of the word in positive and negative documents. We extract a list of words that are polar in the movie domain, but cannot be in a universal sentiment lexicon. Our work is important because without incorporation of such domain specific polar words, the recall of a sentiment analysis system deteriorates. Experimental results show that our proposed method is promising and can be implemented for any domain. Our future work will focus on improving the precision by incorporating the effects of conjunction and negation.

References

Alex Cheng and Oles Zhulyn. 2012. *"A System For Multilingual Sentiment Learning On Large Data Sets"*. Proceedings of the 24th International Conference on Computational Linguistics.

Andrea Esuli and Fabrizio Sebastiani. 2005. *"Determining the semantic orientation of terms through*

gloss classification". Proceedings of the 14th ACM international conference on Information and knowledge management, 617–624.

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 1997. *"Thumbs up?: sentiment classification using machine learning techniques"*. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 79–86.

Eric Breck, Yejin Choi and Claire Cardie. 2007. *"Identifying expressions of opinion in context"*. Proceedings of the 20th international joint conference on Artifical intelligence, 2683–2688.

Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen. 2009. *"Expanding domain sentiment lexicon through double propagation"*. Proceeding of 21st International joint conference on Artificial intelligence, 1199–1204.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. *"Fully automatic lexicon expansion for domain-oriented sentiment analysis"*. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 355–363.

Hong Yu and Vasileios Hatzivassiloglou. 2003. *"Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences"*. Proceedings of the 2003 conference on Empirical methods in natural language processing, 129–136.

Janyce Wiebe. 2000. *"Learning Subjective Adjectives from Corpora"*. Proceedings of the Seventeenth National Conference on Artificial Intelligence, 735–740.

Minqing Hu and Bing Liu. 2004. *"Mining and summarizing customer reviews"*. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168–177.

Peter D Turney. 2002. *"Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews"*. Proceedings of the 40th annual meeting on association for computational linguistics, 417–424.

Theresa Wilson, Janyce Wiebe and Paul Hoffmann. 2005. *"Recognizing contextual polarity in phrase-level sentiment analysis"*. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 347–354.

Vasileios Hatzivassiloglou and Katheleen R McKeown. 1997. *"Recognizing contextual polarity in phrase-level sentiment analysis"*. Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, 174–181.