Augmenting WordNet with Polarity Information on Adjectives

Alekh Agarwal
Deptt. of Comp Science & Engg.
I.I.T. Bombay Mumbai 400076
alekh@cse.iitb.ac.in

Pushpak Bhattacharyya
Deptt. of Comp Science & Engg.
I.I.T. Bombay Mumbai 400076
pb@cse.iitb.ac.in

June 28, 2005

Abstract

Polarity of a word refers to its strength in a classification, typically in a good *vs* bad sense. This paper describes a technique to effectively compute the polarity information for *Adjectives*. We further propose to introduce a new kind of link in the *WordNet* and associate a polarity score with each Adjective in the *WordNet* database. We also show the inter-dependence of subjectivity and polarity of a word. Finally, we demonstrate the need for incorporating such information in *WordNet* by using this information for effective classification of sentences as *subjective* and *objective*.

1 Introduction

Over the years *WordNet* has evolved as a very important and widely used lexical tool. It has been used in a wide range of applications in Natural Language Systems. A few of them are word-sense disambiguation(P. Bhattacharyya et al. GWC 2004), measurement of text similarity(P.Bhattacharyya et al. 2002), text summarization(P.Bhattacharyya et al. LREC 2004) *etc*.

A closely related field that has been studied in a great deal of detail recently is that of Sentiment and Subjectivity Analysis. Typically, and moreover intuitively, adjectives have been found to be very strong indicators of the sentiment of a document. So it is not very surprising that many of the approaches presented for these tasks tend to rely on the notion of the strength of an adjective, and on a more general note, the sentiment contained in a phrase, in various ways. This is especially highlighted in the works of Turney[3], Vegnaduzzo[2] and Baroni et al.[1].

With the increasing use of these techniques need has been felt for the incorporation of polarity and subjectivity information in important lexical databases like *WordNet* (Baroni et al. [1]). This would enable applications to access polarity information in a much faster and convenient way.

In this paper we present a technique to effectively extract polarity related information from the web using an Information Theory based measure. We then propose to enhance the *WordNet* link structure by the addition of *isopolarity* links. These links would be connecting adjectives with a similar degree of polarity. The exact measures of this similarity would be described later on in this paper. We then discuss how useful it is to incorporate the actual polarity scores for every adjective in the *WordNet* database. We also demonstrate how the notions of polarity and subjectivity are closely associated at the word level. Then we go ahead and use this observation to determine the degree of subjectivity of complete sentences using the polarity scores of the adjectives contained in them.

The results of our experiments were very encouraging. Taking the adjectives in *WordNet* with a familiarity count greater than 3, we were able to classify them into 3 clusters. Of these, the two large clusters consisted largely of common positive and negative adjectives. We also observed that words with high subjectivity ratings are typically associated with a higher polarity. We were able to show that using the polarity scores in a supervised learning technique to identify subjective sentences improves the accuracy of classification.

The organization of the rest of the paper is as follows: In section 2 we do a brief review of the past work. In section 3, we describe the various aspects of our approach in detail. In section 4 we describe the exact setup used for our experiments. Our experiments and results are presented in section 5. We conclude the paper by drawing our conclusions and outlining the possible future work in the field in section 6.

2 Previous Work

Polarity and Subjectivity of words has been studied previously in two main kinds of work. One category consists of the works exclusively aiming at the determination of effective metrics for representing the polarity and subjective content of words(Vegnaduzzo, 2004 [2]; Baroni et al. 2004 [1]). Other type consists of the works that do such a study as a part of a larger study on sentiment analysis and related domains which typically have the determination of word polarities as a subtask(Turney, 2002 [3]). Our work lies in the first category but we also go ahead to use the information extracted to perform the task of determining the subjectivity of sentences.

In the initial phases the task of assigning polarity scores has mainly been performed by linguistic experts. However, such a task depends highly on the experts' biases and can typically be done only to the extent of partitioning the words into a few categories rather than assigning ratings to every word on a continuous scale. A landmark theory on assigning polarity scores to adjectives in various kinds of classification tasks is that of Charles Osgood[8]. A technique for effective determination of adjective polarities using this theory and the *WordNet* Synonym graph was presented in Kamps et al.[5]. However, due to the structure of this graph, they could determine these ratings for only about 25% adjectives in the *WordNet* database. Some very important polar adjectives like *excellent* were among the ones that couldn't be reached by their technique. Moreover, this technique fails to generalize for other Parts-of-Speech.

The work closest to ours is perhaps the set of studies conducted by Turney on the use of PMI-IR technique for the mining of polarity scores for adjectives from the web([3],[6]). Our paper largely builds on the approach presented in this paper and then demonstrates how to use this information to augment *WordNet*.

3 Our Method

3.1 Setup

The core of our technique is a measure based on the Point-Wise Mutual Information(PMI) of two words. We used AltaVista Search Engine[1] to measure the PMI between a word and two anchor words. The two anchor words are chosen such that they lie on the opposite ends of the polarity spectrum. Then we used these PMI values for a word to arrive at its polarity score. After this we applied K-Means clustering algo to these scores to obtain the clusters of adjectives on the basis of polarity.

The next step was to take a list of subjective words and obtain the subjectivity scores using a PMI measure again. We went ahead to show the correspondence between polarity scores of words and their subjectivity content. Finally, we moved from word level to sentence level and used the polarity scores of words in conjugation with text classification techniques to distinguish between subjective and objective sentences.

3.2 Determining Polarity scores of Adjectives

We used a Mutual Information based measure for computing the polarity scores of adjectives. The problem of data sparsity for computation of such a statistical measure is always a great bother. However, Turney (2001) [3] introduced the concept of using the Web for determination of these scores using queries made to a Search Engine.

The PointWise Mutual Information between two words is defined as

$$PMI(w1, w2) = log_2 \frac{P(w1, w2)}{P(w1)P(w2)} \tag{1}$$

This is nothing but the ratio of the actual probability two words being seen together to the probability of their being seen together if their occurrence was independent of each other. Hence, thought in an intuitive way, it reflects the degree of association between the two words.

[1] http://www.altavista.com/search/adv

Now let us consider two words of strongly opposite polarity. In our case we took them to be *excellent* and *poor*. Now we define the function Polarity Score for a word $w, PS(w)$ as

$$PS(w) = PMI(w, \textit{excellent}) - PMI(w, \textit{poor}) \quad (2)$$

Now as evident from common sense and pointed out in Baroni et al. [1], words with similar polarity will have a higher rate of co-occurrence. Now consider any positively polar word. As per the above statement, its rate of co-occurrence and hence PMI with the word *excellent* will be higher than that with *poor*. Thus a higher value of $PS(w)$ reflects a higher positive polarity.

To determine the PMI values for the adjectives we used the method of gathering statistics through search engine queries as described by Turney(2002) [3]. In this method $PMI(w1, w2)$ is measured as:

$$PMI(w1, w2) = \frac{Hits(w1ANDw2)}{Hits(w1)Hits(w2)} \quad (3)$$

Here the AND operator is to ask the Search Engine to look for simultaneous occurrences of both $w1$ and $w2$ on the same page. Although, Turney (2002)[3] had used the NEAR operator in his work which is known to give better results than the AND operator([1]), AltaVista has stopped supporting the NEAR operator now. So we had to use the AND operator. These values were retrieved for each of the 21436 adjectives in *WordNet's* database and the Polarity Scores were recorded.

It is important to point out here that although many other statistical measures like Latent Semantic Analysis have also been proposed for this task, PMI has been shown to outperform them([6]). Its ease of calculation is of course an added advantage.

We experimented with some other measures ourselves and tried to take the effect of $Hits(excellent, poor)$, *i.e.* the number of co-occurrences of the two anchor words into account. However, the results with PMI scores were much better than these measures.

3.3 Incorporation of Polarity Information in WordNet

Once we had the polarity scores for the adjectives, we ran a k-means clustering algorithm with the Calinski and Harbasz's stopping rule([7]) to select the optimum number of clusters. In this rule,we use a Variance Ratio Criterion (VRC) defined as:

$$VRC = \frac{\frac{BGSS}{k-1}}{\frac{WGSS}{n-k}} \quad (4)$$

where,

$WGSS$ is the Within-Cluster Sum of Squared Distances about the Centroids,
$BGSS$ is the total between cluster Sum of Squared Distances,
k is the number of clusters, and
n is the number of data points.

These clusters contain words having similar polarities. Hence we propose to link these words with a new kind of link called the *isopolarity* link in *WordNet*. We also propose to store with each word, its polarity score calculated in the previous section.

3.4 Detection of Subjective Content of Adjectives

For this task, we employed the technique described in Baroni et al.(2004) [1]. We took the list of 35 seeds that Baroni et al.[1] and Vegnaduzzo[2] had used earlier and randomly picked 10 adjectives out of it. We then calculated the PMI scores for each Adjective in *WordNet's* database that had a familiarity count greater than 3 in *WordNet's* tagged corpus.

We calculated these PMI scores with each of the 10 seed words. The formula used for PMI computation was a little different in this case. We used the following expression:

$$PMI(w1, w2) = log_2 N \frac{Hits(w1ANDw2)}{Hits(w1)Hits(w2)} \quad (5)$$

The extra factor of N was in the numerator was taken as 1 billion, the approximate number of documents indexed by AltaVista. This expression comes from taking the Maximum Likelihood Estimators for the probabilities. Although the constant factor is not of great importance as it is the relative scores that matter, the values with $N = 1$ were too small to be worked with.

Then we took the *first quartile* of these scores(which had given the best results in Baroni's study[1]) and recorded it as the subjectivity rating of the word. These scores were used to observe the correspondence between polarity and subjectivity ratings for adjectives.

3.5 Using the Adjective Polarity values to compute Sentence Subjectivity

To demonstrate how the polarity scores can be used in practical applications, we used these scores to determine the subjectivity of sentences. An SVM based classifier formed the core of this approach. We used bag-of-words features. The top 1000 adjectives and 5000 non-adjective words were chosen as features on the basis of their occurrence in the dataset. With the non-adjectival features binary values were used in the feature vectors. For adjectival features, we tried two different approaches. One was to use the weight of adjective if the adjective was found in the document, 0 otherwise. In the second case we used binary values just like other features. The five-fold cross validation accuracies were recorded for both the cases.

4 Evaluation

4.1 Experimental Setup

We used the *wget* program on linux to retrieve the webpages corresponding to queries generated for words and then extracted the hits from them. The WordNet::QueryData module[2] by Jason Rennie was used for querying WordNet for information like list of adjectives and familiarity counts of adjectives. The set of seeds for the subjectivity score determination for adjectives was derived from the list of 35 seeds used by Baroni et al.(2004) [1] and Vegnaduzzo(2004) [2].

The sentence corpus used for Subjective *vs* Objective classification was the subjectivity corpus introduced by Bo Pang et al. in ACL 2004[4]. [3]. This corpus contains 5000 plot summaries and 5000 review snippets that were collected from *www.rottentomatoes.com*. The entire corpus was run through a POS tagger as the POS tags were needed for later tasks. The tagger used was Stanford Log-Linear Model Tagger v1.0[4]. The resulting sentences were used for subjectivity detection.

The SVM package used was libsvm-2.71[9]. The clustering algo used for determining *isopolar* groups was k-means clustering. We used a C Clustering Library implemented by M.J.L. de Hoon et al.[10] for a basic implementation of k-means clustering and added the Calinski and Harbasz's stopping rule to it for determining the initially unknown number of clusters.

5 Experimental Results

The first thing we did was to retrieve a list of all the adjectives in *WordNet's* lexical database. With each of the adjectives we performed queries to measure their PMI scores with respect to *excellent* and *poor*. Thus a total of 42874 queries was performed in this step.

When the Polarity scores obtained through this process were analyzed, we realized that the words with negative scores had greatly outnumbered those with positive scores. Also the words with very large scores on either side were mostly obscure words. On performing k-means clustering on this, we obtained the optimum number of clusters as 3. But the words were largely cluttered up into two clusters, both with their centroids on the negative side. We further tried an Expectation Maximization clustering on these scores. Here cross-validation resulted yielded 7 clusters as the optimum value. However, words were still mainly cluttered in two clusters, both with negative mean and together accounting for over 90% adjectives in the *WordNet's* database.

We realized that the scores for the obscure words were not very good indicators of their polarities as their co-occurrence with one of the anchor word was usually too low leading to an abnormally high score on either end of the spectrum. This incorrect readings presented difficulties for the clustering algo downstream by adding a huge number of noisy points.

To take care of the above problem, we extracted the adjectives in the *WordNet's* database which had a familiarity count greater than 3. This gave 745 words. A cursory glance was enough to realize that all the obscure noise words had been filtered by this step.

[2] http://people.csail.mit.edu/ jrennie/WordNet

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

[4] http://www-nlp.stanford.edu/software/tagger.shtml

We went ahead to extract the polarity scores for these words. With the noise words removed, the scores were much easier to analyze and now a clear correlation between polarities and scores could be seen. We went ahead to apply the k-means clustering on this set. Of these there were two large clusters with 493 and 236 words. The third cluster was a small one with just 16 words. Except the anchor word *poor* none of the words in the cluster had a high degree of polarity but had large negative ratings.

However, the more important clusters are the other two. The 436 word cluster has majority of the words positive polarity. (Note that by majority of the words having a positive polarity, here we mean that the majority of the words actually having a positive polarity. On the basis of scores, this cluster has all the words with positive scores. But some positive polarity words end up with negative scores which is a classification error that'll be always present in any unsupervised learning method).

The 236 word cluster contains most of the words with negative polarity. The cluster id and polarity scores for some of the common adjectives have been listed in Table 1.

In the table we can clearly see how the words having their polarity scores very close have a close association among their polarities. Although, it maybe pointed out that the words in a cluster vary over a wide range of polarities, but this is due to the wide range of polarity scores contained in clusters. If smaller clusters are desired, all one has to do is to increase the number of clusters in the clustering step.

One important observation that we would like to point out here is that the polarity scores range from -3.0 to 1.34. As the ranges on either side of zero are not uniform, further works can try to use supervised learning to learn an optimum threshold score to label words as positively and negatively polar. For such a task a small list of adjectives with hand-assigned labels would be needed from which threshold can be learned.

We would also like to describe some of the other measures we tried out to measure polarity scores. The basic notion in our mind was that $Hits(a1, a2)$ where $a1$ and $a2$ are the two anchor words represent the association between the two anchor words. So we should take this value into account. We tried to do this by trying out the func-

	Word	Polarity Score	Cluster
1.	comparative	-3.00	0
2.	poor	-2.91	0
3.	legislative	-2.81	0
4.	democratic	-2.51	0
5.	political	-1.77	0
6.	slow	-0.74	1
7.	passive	-0.72	1
8.	late	-0.71	1
9.	monotnous	-0.67	1
10.	highest	-0.15	1
11.	accurate	-0.14	1
12.	effi ecient	-0.13	1
13.	good	-0.12	1
14.	outstanding	0.56	1
15.	unique	0.57	1
16.	spectacular	0.59	1
17.	sporting	1.34	1
18.	raw	-1.58	2
19.	crude	-1.36	2
20.	cruel	-1.35	2
21.	furious	-0.99	2
22.	fi erce	-0.99	2
23.	bad	-0.98	2
24.	dangerous	-0.96	2
25.	scattered	-0.79	2
26.	uneven	-0.78	2
27.	lesser	-0.78	2
28.	lacking	-0.77	2
29.	irregular	-0.76	2
30.	lost	-0.76	2

Table 1: Polarity Scores and Clusters For a few Common Words

tions

$$f_1(w) = \frac{\frac{Hits(wANDexcellent)}{Hits(excellent)} - \frac{Hits(wANDpoor)}{Hits(poor)}}{Hits(excellent, poor)} \qquad (6)$$

and

$$f_2(w) = N \frac{\frac{Hits(wANDexcellent)}{Hits(excellent)Hits(w)} - \frac{Hits(wANDpoor)}{Hits(poor)Hits(w)}}{Hits(excellent, poor)} \qquad (7)$$

However, the values in both cases didn't have a good correlation with word polarity and the results for clustering were also bad in these cases. As for statistically sophisticated measures, studies in the past have already proved that PMI is a better performer([6],[1]).

After this we took up the task of analyzing the correlation between word polarity and subjectivity scores. The subjectivity scores were evaluated using the technique described in Baroni et al.(2004)[1]. The details are as given in section 3.3. The results were encouraging. Of the top 25 subjective words, only 8 had the polarity scores lying in the range of -0.5 to 0.5. This number was 19 in the first 50 words. 36 such words were present in the top 100 subjective words. Please note that the values -0.5 and 0.5 were handpicked and better analysis can be done if the optimum threshold for polarity determination is learned using supervised learning as pointed out earlier.

The final task we took up was to see the effect of polarity scores on determination of subjectivity of sentences using the approach described in section 3.4. The accuracy using binary values in the feature vector for both adjectival and non-adjectival features was found out to be 55.93%. On using polarity scores instead of a binary value for adjectival features, the accuracy rose to 61.1%.

6 Conclusions

The incorporation of polarity information in *WordNet* is definitely greatly needed taking into account the increasing number of works that rely on such information. Our approach shows that this is definitely feasible using automated techniques. Small misclassifications do occur and human intervention would be definitely needed at times if a very high precision is desired. However, our work largely suffices to provide the basic infrastructure to do such a task.

One may point out here that a drawback of our approach was the need to cut the number of adjectives down to 745 for eliminating noise words. We would like to point out here, that as far as the individual scores are concerned, this approach can fetch them even for the words that we have eliminated. The real trouble is caused at the clustering step. So a possible option is to keep the individual scores for these words in *WordNet* but to omit the *isopolarity* links for them. At this point, one more thing that we would like to point out is that the value of familiarity count greater than 3 was chosen randomly and the scope of clustering can be increased. Real problem is due to the words that have familiarity counts of 0. And also, as we have been saying from the beginning, the whole aim of this work is to make the polarity data available for practical applications that heavily rely on such information. So when we discuss the omission of *isopolarity* links for words with a low familiarity count, we are not really affecting the users of the polarity information in the sense that such words are not very common. The strength of our approach is that the common and frequent words will always be reliably linked with a fairly high degree of accuracy.

Here, we would also like to contrast our results against those of Kamps et al. [5], which had tried to show the association between *WordNet* graph and polarity scores of adjectives. We pointed out earlier that their approach is not able to find the polarity scores for all words. A casual observer might comment that this drawback is there in our method also. However, there is a very important difference. Their approach relied on the *WordNet* synonym link structure. So if a word is not connected to the anchor word through synonymy links, then its score cant be evaluated. This leads to some very common adjectives like *excellent* being left out using *good* and *bad* as the two anchor words. Our approach can always calculate the polarity scores for any word and the reliability of clustering increases for common words. So the case of important words that might be frequently needed by applications being left out in our approach would hardly ever arise. Another important distinction is that our technique is not specific to adjectives or even unigrams for that matter.

Also the dependence between subjectivity and polarity that is reflected in our work provides and interesting line of work in future. The gain in accuracy of classification on using polarity scores is also heartening and better ways of incorporating this information in the classification technique is bound to improve the results.

Future work A very interesting line for the future works would be try and obtain the polarity information separately for different senses of a word. This would require using the facts that there should be a very high correlation between polarity scores of the words in a *Synset*. *WordNet* glosses for the synsets can also be used to determine these scores.

Also other techniques to obtain the *isopolar* groups from the polarity scores can be experimented with. A very big drawback is the withdrawl of NEAR operator AltaVista. Looking into other techniques and sources to obtain this information can also been looked into.

Also we have pointed out earlier that polarity scores are not evenly distributed about 0. So looking into appropriate supervised learning techniques to determine an appropriate threshold can be considered.

The dependence between subjectivity and polarity can also be used in strengthening the data for one if the data for the other is available.

7 Acknowledgments

We would like to thank Marco Baroni for providing us with the tagged list of adjectives and set of seeds used in his work.

References

[1] M. Baroni and S. Vegnaduzzo, *Identifying Subjective Adjectives through Web-based Mutual Information,* In Ernst Buchberger (ed.), Proceedings of KONVENS 2004, Vienna: GAI. 17-24.

[2] S. Vegnaduzzo, 2004, *Acquisition of subjective adjectives with limited resources,* In AAAI Spring Symposium Technical Report: Exploring Affect and Attitude in Text.

[3] Peter Turney, *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classication of reviews*, In Proc. of the ACL. 2002.

[4] Bo Pang and Lillian Lee, *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts*, Proceedings of ACL, 2004.

[5] Jaap Kamps, Robert J. Mokken, Maarten Marx, and Maarten de Rijke. *Using WordNet to measure semantic orientation of adjectives.* In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), volume IV, pages 1115-1118. European Language Resources Association, Paris, 2004.

[6] Turney, P.D. 2001. *Mining the Web for synonyms: PMI-IR versus LSA on TOEFL*. Proceedings of the Twelfth European Conference on Machine Learning (pp. 491-502). Berlin: Springer-Verlag.

[7] Calinski & Harabasz 1974 *A Dendrite Method for Cluster Analysis*. Communications in Statistics, 3(1), (1974) 127.

[8] Osgood, C. E., G. J. Succi, and P. H. Tannenbaum, 1957. *The Measurement of Meaning.* University of Illinois Press, Urbana IL.

[9] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001.Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

[10] M.J.L. de Hoon, S. Imoto, J. Nolan, and S. Miyano, *Open Source Clustering Software,* Bioinformatics 20(9): 1453-1454 (2004).