Using Pivot-Based Paraphrasing and Sentiment Profiles to Improve a Subjectivity Lexicon for Essay Data

Beata Beigman Klebanov, Nitin Madnani, Jill Burstein
Educational Testing Service
660 Rosedale Road, Princeton, NJ 08541, USA
{bbeigmanklebanov,nmadnani,jburstein@ets.org}

Abstract

We demonstrate a method of improving a seed sentiment lexicon developed on essay data by using a pivot-based paraphrasing system for lexical expansion coupled with sentiment profile enrichment using crowdsourcing. Profile enrichment alone yields up to 15% improvement in the accuracy of the seed lexicon on 3-way sentence-level sentiment polarity classification of essay data. Using lexical expansion in addition to sentiment profiles provides a further 7% improvement in performance. Additional experiments show that the proposed method is also effective with other subjectivity lexicons and in a different domain of application (product reviews).

1 Introduction

In almost any sub-field of computational linguistics, creation of working systems starts with an investment in manually-generated or manually-annotated data for computational exploration. In subjectivity and sentiment analysis, annotation of training and testing data and construction of subjectivity lexicons have been the loci of costly labor investment.

Many subjectivity lexicons are mentioned in the literature. The two large manually-built lexicons for English – the General Inquirer (Stone et al., 1966) and the lexicon provided with the OpinionFinder distribution (Wiebe and Riloff, 2005) – are available for research and education only[1] and under GNU GPL license that disallows their incorporation into proprietary materials,[2] respectively. Those wishing to integrate sentiment analysis into products, along with those studying subjectivity in languages other than English, or for specific domains such as finance, or for particular genres such as MySpace comments, reported construction of lexicons (Taboada et al., 2011; Loughran and McDonald, 2011; Thelwall et al., 2010; Rao and Ravichandran, 2009; Jijkoun and Hofmann, 2009; Pitel and Grefenstette, 2008; Mihalcea et al., 2007).

In this paper, we address the step of expanding a small-scale, manually-built subjectivity lexicon (a seed lexicon, typically for a domain or language in question) into a much larger but noisier lexicon using an automatic procedure. We present a novel expansion method using a state-of-the-art paraphrasing system. The expansion yields a 4-fold increase in lexicon size; yet, the expansion alone is insufficient in order to improve performance on sentence-level sentiment polarity classification.

In this paper we test the following hypothesis. We suggest that the effectiveness of the expansion is hampered by (1) introduction of opposite-polarity items, such as introducing *resolute* as an expansion of *forceful*, or *remarkable* as an expansion of *peculiar*; (2) introduction of weakly polar, neutral, or ambiguous words as expansions of polar seed words, such as generating *concern* as an expansion of *anxiety* or *future* as an expansion of *aftermath*;[3] (3) inability to distinguish between stronger or clear-cut versus weaker or ambiguous sentiment and to make a differential use of those.

We address items (1) and (2) by enriching the lexicon with sentiment profiles (section 3), and propose

[1] http://www.wjh.harvard.edu/ inquirer/j1_1/manual/

[2] http://www.gnu.org/copyleft/gpl.html

[3] Table 2 and Figure 1 provide support to these assessments.

Transactions of the Association for Computational Linguistics, 1 (2013) 99–110. Action Editor: Chris Callison-Burch.
Submitted 12/2012; Published 5/2013. ©2013 Association for Computational Linguistics.

a way of effectively utilizing this information for the sentence-level sentiment polarity classification task (sections 5 and 6). Profile-enrichment alone yields up to 15% increase in performance for the seed lexicon when using different machine learning algorithms; paraphraser-based expansion with sentiment profiles improves performance by an additional 7%. Overall, we observe an improvement of up to 25% in classification accuracy over the seed lexicon without profiles.

In section 7, we present comparative evaluations, demonstrating the competitiveness of the expanded and profile-enriched lexicon, as well as the effectiveness of the expansion and enrichment paradigm presented here for different subjectivity lexicons, different lexical expansion methods, and in a different domain of application (product reviews).

2 Building Subjectivity Lexicons

The goal of our sentiment analysis project is to allow for the identification of sentiment in sentences that appear in essay responses to a variety of tasks designed to test English proficiency in both native- and non-native-speaker populations in a standardized assessment as well as in an instructional settings. In order to allow for the future use of the sentiment analyzer in a proprietory product and to ensure its fit to the test-taker essay domain, we began our work with the construction of a seed lexicon relying on our materials (section 2.1). We then used a statistical paraphrasing system to expand the seed lexicon (section 2.2).

2.1 Seed Lexicon

In order to inform the process of lexicon construction, we randomly sampled 5,000 essays from a corpus of about 100,000 essays containing writing samples across many topics. Essays were responses to several different writing assignments, including graduate school entrance exams, non-native English speaker proficiency exams, and professional licensure exams. Our seed lexicon is a combination of (1) positive and negative sentiment words manually selected from a full list of word types in these data, and (2) words marked in a small-scale annotation of a sample of sentences from these data for all positive and negative words. A more detailed description of the construction of seed lexicon can be found in Beigman Klebanov et al (2012). The seed lexicon contains 749 single words, 406 positive and 343 negative.

2.2 Expanded Lexicon

We used a pivot-based lexical and phrasal paraphrase generation system (Madnani and Dorr, 2013). The paraphraser implements the pivot-based method as described by Bannard and Callison-Burch (2005) with several additional filtering mechanisms to increase the precision of the extracted pairs. The pivot-based method utilizes the inherent monolingual semantic knowledge from bilingual corpora: We first identify phrasal correspondences between English and a given foreign language F, then map from English to English by following translation units from English to the other language and back. For example, if the two English phrases $e1$ and $e2$ both correspond to the same foreign phrase f, then they may be considered to be paraphrases of each other with the following probability:

$$p(e1|e2) \approx p(e1|f)p(f|e2)$$

If there are several pivot phrases that link the two English phrases, then they are all used in computing the probability:

$$p(e1|e2) \approx \sum_{f'} p(e1|f')p(f'|e2)$$

Seed	Expansion	Seed	Expansion
abuse	exploitation	costly	onerous
accuse	reproach	dangerous	unsafe
anxiety	disquiet	improve	reinforce
conflict	crisis	invaluable	precious

Table 1: Examples of paraphraser expansions.

Some examples of expansions generated by the paraphraser are shown in Table 1. More details about this kind of approach can be found in Bannard and Callison-Burch (2005). We use the French-English parallel corpus (approximately 1.2 million sentences) from the corpus of European parliamentary proceedings (Koehn, 2005) as the data on which pivoting is performed to extract the paraphrases. However, the base paraphrase system is susceptible to large amounts of noise due to the imperfect bilingual word alignments. Therefore, we implement additional heuristics in order to minimize the number of noisy paraphrase pairs (Madnani and Dorr, 2013). For example, one such heuristic filters out pairs where a function word may have been inferred as a paraphrase of a content word. For the lexicon expansion experiment reported here, we use the top 15 single-word paraphrases for every word from the seed lexicon, excluding morphological variants of the seed word. This process results in an expanded lexicon of 2,994 different words, 1,666 positive and 1,761 negative (433 words are in both the positive and the negative lists). The expanded lexicon includes the seed lexicon.

3 Inducing sentiment profiles

Let γ^w be the sentiment profile of the word w.

$$\gamma^w = (p_w^{pos}, p_w^{neg}, p_w^{neu}) \quad (1)$$

where $\Sigma_{i \in \{pos,neg,neu\}} \; p_w^i = 1$. Thus, a sentiment profile of a word is essentially a 3-sided coin, corresponding to its probability of coming out positive, negative, and neutral, respectively.

3.1 Estimating sentiment profiles

Our goal is to estimate the profile using outcomes of multiple trials as follows. For every word, a person is shown the word and asked whether it is positive, negative, or neutral. A person's decision is modeled as flipping the coin corresponding to the word, and recording the outcome – positive, negative, or neutral. We run N=20 such trials for every word in the expanded lexicon using the CrowdFlower crowdsourcing site,[4] for a total cost of $800. We use maximum likelihood estimate of sentiment profile:

$$\hat{p}_w^i = n_w^i \quad (2)$$

where n_w^i is the proportion of N trials on the word w that fell in cell $i \in \{pos, neg, neu\}$. Table 2 shows some estimated profiles.

Following Goodman (1965) and Quesenberry and Hurst (1964), we calculate confidence intervals for the parameters p_w^i:

$$(\hat{p}_w^i)^- = (B + 2n_w^i - T)/(2(N + B)) \quad (3)$$

Word	$\hat{p}_w^{pos}$	$\hat{p}_w^{neu}$	$\hat{p}_w^{neg}$
forceful	0	0.15	0.85
resolute	0.8	0.15	0.05
peculiar	0.05	0.15	0.8
remarkable	1	0	0
anxiety	0	0	1
concern	0.25	0.4	0.35
absurd	0	0	1
laughable	0.5	0.05	0.45
deadly	0	0	1
fateful	0.25	0.45	0.3
consequence	0.05	0.15	0.8
outcome	0.15	0.85	0

Table 2: Examples of estimated sentiment profiles. Words in gray are expansions generated from words in the preceding row; note the difference in the profiles.

$$(\hat{p}_w^i)^+ = (B + 2n_w^i + T)/(2(N + B)) \quad (4)$$

where

$$T = \sqrt{B[B + 4n_w^i(N - n_w^i)/N])} \quad (5)$$

For confidence α that all p_w^i, $i \in \{pos, neg, neu\}$ are simultaneously within their respective intervals, the value of B is determined as the upper $\alpha/3 \times 100^{th}$ percentile of the χ^2 distribution with one degree of freedom. We use α=0.1, resulting in B=4.55. The resulting interval is about 0.2 around the estimated value when $\hat{p}_w^i$ is close to 0.5, and somewhat narrower for $\hat{p}_w^i$ closer to 0 or 1. We will use this information when inducing features from the profiles.

3.2 Sentiment distributions of the lexicons

The estimated sentiment profiles per word allow us to visualize the distributions of the two lexicons. In Figure 1, we plot the number of entries in the lexicon as a function of the difference in positive and negative parts of the profile, in 0.2-wide bins. Thus, a word w would be in the second-leftmost bin if $-0.8 < (\hat{p}_w^{pos} - \hat{p}_w^{neg}) < -0.6$.

While the expansion process more than doubles the number of words in the highest bins for both the positive and the negative polarity, it clearly introduces a large number of words in the low- and medium bins into the lexicon. It is in this sense that the expansion process is noisy; apparently, seed words with clear and strong polarity

[4] www.crowdflower.com

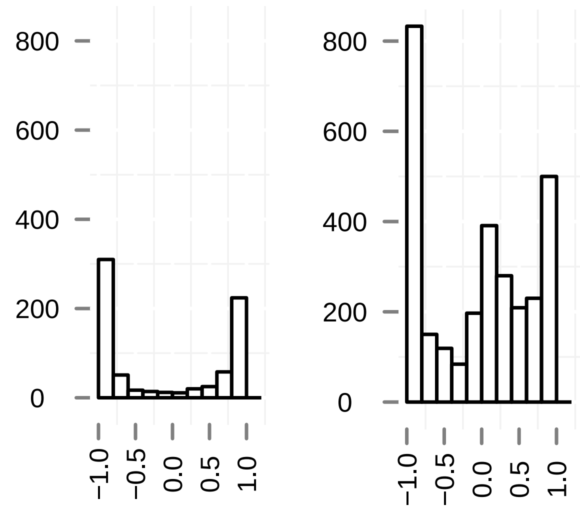

Figure 1: Sentiment distributions for the seed (left) and the expanded (right) lexicons.

are often expanded into low intensity, neutral, or ambiguous ones, as in pairs like *absurd/laughable*, *deadly/fateful*, *anxiety/concern* shown in Table 2.

4 Related Work

The most popular seed expansion methods discussed in the literature are based on WordNet (Miller, 1995) or another lexicographic resource, on distributional similarity with the seeds, or on a mixture thereof (Cruz et al., 2011; Baccianella et al., 2010; Velikovich et al., 2010; Qiu et al., 2009; Mohammad et al., 2009; Esuli and Sebastiani, 2006; Kim and Hovy, 2004; Andreevskaia and Bergler, 2006; Hu and Liu, 2004; Kanayama and Nasukawa, 2006; Strapparava and Valitutti, 2004; Kamps et al., 2004; Takamura et al., 2005; Turney and Littman, 2003; Hatzivassiloglou and McKeown, 1997). The paraphrase-based expansion method is in the distributional similarity camp; we also experimented with WordNet-based expansion as descibed in section 7.2.

The task of assigning sentiment profiles to words in a sentiment lexicon has been addressed in the literature. SentiWordNet assigns profiles to all words in WordNet based on a propagation algorithm from a small seed set manually annotated by a small number of judges (Baccianella et al., 2010; Cerini et al., 2007). Andreevskaia and Bergler (2006) use graph propagation algorithms on WordNet to assign centrality scores in positive and negative categories; a similar approach based on web-scale co-occurrence graphs is discussed in Velikovich et al (2010). Thelwall et al (2010) manually annotated a set of words for strength of sentiment and used machine learning to fine-tune it. Taboada et al (2011) produced an expert annotation of their lexicon with strength of sentiment. Subasic and Huettner (2001) manually built an affect lexicon with intensities. Wiebe and Riloff (2005) classifed lexicon entries into weakly and strongly subjective, based on their relative frequency of appearance in subjective versus objective contexts in a large annotated dataset.

Our sentiment profiles are best thought of as relatively fine-grained *priors* for the sentiment expressed by a given word out-of-context. These reflect a mixture of strength of sentiment ($\hat{p}^{pos}_{good} > \hat{p}^{pos}_{decent}$), contextual ambiguity (*concern* can be interpreted as similar to *worry* or to *care*, as in “Her condition was causing concern” versus “He showed genuine concern for her”), and dominance of a polar connotation (*abandon* is $\hat{p}^{neg}$=1; it has a negative overtone even if the actual sense is not that of *desert* but of *vacate*, as in “You must abandon your office”).

To the best of our knowledge, this paper presents the first attempt to integrate judgements obtained through crowdsourcing on a large scale into a sentiment lexicon, showing the effectiveness of this lexicon-enrichment procedure for a sentiment classification task.

5 Using profiles for sentence-level sentiment polarity classification

To evaluate the usefulness of the lexicons, we use them to generate features for machine learning systems, and compare performance on 3-way sentence-level sentiment polarity classification. To ensure robustness of the observed trends, we experiment with a number of machine learning algorithms: SVM Linear and RBF, Naïve Bayes, Logistic Regression (using WEKA (Hall et al., 2009)), and c5.0 Decision Trees (Quinlan, 1993).[5]

5.1 Data

We generated the data for training and testing the machine learning systems as follows. We used our

[5] available from http://rulequest.com/

pool of 100,000 essays to sample a second, non-overlapping set of 5,000 essays, so that no essay used for lexicon development appears in this set. From these essays, we randomly sampled 550 sentences, and submitted them to sentiment polarity annotation by two experienced research assistants; 50 double-annotated sentenced showed κ=0.8. TEST set contains the 43 agreed double-annotated sentences, and additional 238 sampled from the 500 single-annotated sentences, 281 sentence in total. The category distribution in the TEST set is 46.6% neutral, 32.4% positive, and 21% negative.

The TRAIN set contains the remaining sentences, plus positive, negative, and neutral sentences annotated during lexicon development, for the total of 1,631 sentences. The category distribution in TRAIN is 39% neutral, 35% positive, 26% negative.

5.2 From lexicons to features

Our goal is to evaluate the impact of sentiment profiles on sentence-level sentiment polarity classification for the seed and the expanded lexicons, while also looking for the most effective ways to represent this information for machine learners.

We implement two baseline systems. One provides the machine learner with the most detailed information contained in a lexicon: **BL-full** has 2 features for every lexicon word, taking the values (1,0) for positive match in a sentence, (0,1) – for negative, (1,1) for a word in both positive and negative parts of the lexicon, and (0,0) otherwise.

The second baseline provides the machine learner with only summary information about the overall sentiment of the sentence. **BL-sum** uses only 2 features: (1) the total count of positive words in the sentence; (2) the total count of negative words in the sentence, according to the given lexicon.

For the sentiment-enriched runs, we construct a number of representations: Int-full, Int-sum, Int-bin, and Int-c. Int-full and Int-sum are parallel to the respective baseline systems. **Int-full** represents each lexicon word as 2 features corresponding to the word's estimated $\hat{p}_w^{pos}$ and $\hat{p}_w^{neg}$, providing the most detailed information to the machine learner. In the **Int-sum** condition, we use $\hat{p}_w^{pos}$ and $\hat{p}_w^{neg}$ for every word to induce 2 features: (1) the sum of positive probabilities of all words in the sentence; (2) the sum of negative probabilities for all words in the sentence, according to the given lexicon.

For **Int-bin** runs, we use bins of the size of 0.2 – half of the maximal confidence interval – to group together words with close estimates. We produce 10 features. For positive bins, the 5 features count the number of words in the sentence that fall in bin$_i$, $1 \leq i \leq 5$, respectively, that is, words with $0.2(i-1) < \hat{p}_w^{pos} \leq 0.2i$. Bin 1 also includes words with $\hat{p}_w^{pos}$ = 0, since these cannot be distinguished with high confidence from $\hat{p}_w^{pos}$=0.1. Note that we do not provide a scale, we merely represent different ranges with different features. This should allow the machine learners the flexibility to weight the different bins differently when inducing classifiers.

The **Int-c** condition represents a coarse-grained setting. We produce 4 features, two for each polarity: (1) the number of words such that $0 \leq \hat{p}_w^{pos} < 0.4$; (2) the number of words such that $0.4 \leq \hat{p}_w^{pos} \leq 1$; similarity for the negative polarity.

Table 3 summarizes conditions and features.

Cond.	#F	Feature Description
BL-full	2|L|	$(\mathbf{1}_{L^{pos} \cap S}(w), \mathbf{1}_{L^{neg} \cap S}(w))$

BL-sum	2	f$_1$=$\|\{w : w \in L^{pos} \cap S\}\|$
		f$_2$=$\|\{w : w \in L^{neg} \cap S\}\|$

Int-full	2|L|	$(\hat{p}_w^{pos}, \hat{p}_w^{neg})\ \forall w \in A$
Int-sum	2	$(\Sigma_{w \in A}\ \hat{p}_w^{pos}, \Sigma_{w \in A}\ \hat{p}_w^{neg})$
Int-bin	10	f$_1$=$\|\{w \in A : 0 \leq \hat{p}_w^{pos} \leq 0.2\}\|$ ... f$_{10}$=$\|\{w \in A : 0.8 < \hat{p}_w^{neg} \leq 1\}\|$

Int-c	4	f$_1$=$\|\{w \in A : 0 \leq \hat{p}_w^{pos} < 0.4\}\|$
		...
		f$_4$=$\|\{w \in A : 0.4 \leq \hat{p}_w^{neg} \leq 1\}\|$

Table 3: Description of conditions. Column 2 shows the number of features. In column 3: **1** is an indicator function; L is a lexicon; L^{pos} is the part of the lexicon containing positive words (same with negatives); S is a sentence for which a feature vector is built; $A = L \cap S$. For all $w \in L - S$ in the -full conditions, w is represented with (0,0).

6 Results

Table 4 shows classification accuracies for 5 machine learning systems across 6 conditions, for the seed and the expanded lexicons.

Let **BL** denote the best-performing baseline (BL-

Machine Learner	**Condition**	**Seed**	**Expanded**
–	Majority	0.466	0.466
c5.0	BL-full	0.441	0.498
	BL-sum	0.512	0.480
	Int-full	0.441	0.498
	Int-sum	0.566	0.616
	Int-bin	0.587	0.641
	Int-c	0.530	0.577
SVM RBF	BL-full	0.466	0.466
	BL-sum	0.527	0.495
	Int-full	0.466	0.466
	Int-sum	0.548	0.601
	Int-bin	0.573	**0.644**
	Int-c	0.530	0.562
SVM Linear	BL-full	0.584	0.566
	BL-sum	0.509	0.502
	Int-full	0.580	0.609
	Int-sum	0.601	0.580
	Int-bin	0.573	0.630
	Int-c	0.569	0.569
Logistic Regression	BL-full	0.545	0.509
	BL-sum	0.545	0.509
	Int-full	0.534	0.502
	Int-sum	0.555	0.584
	Int-bin	0.584	0.616
	Int-c	0.545	0.577
Naïve Bayes	BL-full	0.598	0.584
	BL-sum	0.509	0.473
	Int-full	0.598	0.580
	Int-sum	0.545	0.605
	Int-bin	0.559	0.626
	Int-c	0.537	0.601

Table 4: Classification accuracies on TEST set. Majority baseline corresponds to classifying all sentences as neutral. The best performance is boldfaced. Let BL stand for the best-performing baseline (BL-full or BL-sum) for a combination of machine learner and lexicon. We use Wilcoxon Signed-Rank test, reporting the number of signed ranks (N) and the sum of signed ranks (W). Statistically significant results at p=0.05 are: Int-sum > BL (N=10, W=43); Int-bin > BL (N=10, W=48); Int-bin > Int-sum (N=10, W=43); Int-bin > Int-full (N=10, W=47); Int-sum > Int-full (N=10, W=37); Int-bin > Int-c (N=10, W=55); Int-sum > Int-c (N=10, W=55); Expanded > Seed under Int condition (includes Int-full, Int-sum, Int-bin, Int-c) (N=18, W=152, z=3.3). Differences between Int-full, Int-c, and BL are not significant.

full or BL-sum) for a combination of machine learner and lexicon. The results show that (1) Int-bin > Int-sum > BL = Int-c = Int-full; (2) Expanded > Seed under Int condition. All inequalities are statistically significant at p=0.05 (see caption of Table 4 for details).

First, both the seed and the expanded lexicons benefit from profile enrichment, although, as predicted, the expanded lexicon yields larger gains due to its more varied profiles: The seed lexicon gains up to 15% in accuracy (c5.0 BL-sum vs Int-bin), while the expanded lexicon gains up to 30%, as SVM RBF scores go up from 0.495 to 0.644.

Second, observe that profiling allows the expanded lexicon to leverage its improved coverage: While it is inferior to the best baseline run with the seed lexicon for all systems, it succeeds in improving the seed lexicon accuracies by 5%-12% across the different systems for the Int-bin runs. The best run of the expanded lexicon (Int-bin for SVM RBF) improves upon the best run of the seed lexicon (Int-sum for SVM-linear) by 7%, demonstrating the success of the paraphraser-based expansion once profiles are taken into account. Overall, comparing the best baseline for the seed lexicon with Int-bin condition of the expanded lexicon, we observe an improvement between 5% (0.598 to 0.626 for Naïve Bayes) and 25% (0.512 to 0.641 for c5.0), proving the effectiviness of the paraphrase-based expansion with profile enrichment paradigm.

Third, representing profiles using 10 bins (Int-bin) provides a small but consistent improvement over the summary representation (Int-sum) that sums positivity and negativity of the sentiment-bearing words in a sentence, over a coarse-grained representation (Int-c), as well as over the full-information representation (Int-full). Even Naïve Bayes and SVM linear, known to work well with large feature sets, show better performance in the Int-bin condition for the expanded lexicon. The results indicate that an intermediate degree of detail – between summary-only and coarse-grained representation on the one hand and full-information representation on the other – is the best choice in our setting.

7 Comparative Evaluations

In this section, we present comparative evaluations of the work presented in this paper with respect to related work. This section shows that the paraphrase expansion+profile enrichment solution proposed in this paper is effective for our task beyond off-the-shelf solutions, and that its effectiveness generalizes to sentiment analysis in a different domain. We also show that profile enrichment can be effectively coupled with other methods of lexical expansion, although the paraphraser-based expansion receives a larger boost in performance from profile enrichment than the alternative expansion methods we consider.

In section 7.1, we demonstrate that the paraphrase-based expansion and profile enrichment yield superior performance on our data relative to state-of-art subjectivity lexicons – OpinionFinder, General Inquirer, and SentiWordNet. In section 7.2, we show that profile enrichment can be effectively coupled with other methods of lexical expansion, such as a WordNet-based expansion and an expansion that utilizes Lin's distributional thesaurus. However, we find that the paraphraser-based expansion benefits the most from profile enrichment, and attains better performance on our data than the alterantive expansion methods. In section 7.3, we show that the paraphrase-based expansion and profile enrichment paradigm is effective for other subjecitivy lexicons on other data. We use a dataset of product reviews annotated for sentence-level positivity and negativity as new data for evaluation (Hu and Liu, 2004). We use subsets of OpinionFinder, General Inquirer, and sentiment lexicon from Hu and Liu (2004). We demonstrate that paraphrase-based expansion and profile enrichment improve the accuracy of sentiment classification of product reviews *for every lexicon and machine learner combination*; the magnitude of improvement is 5% on average.

7.1 Competitiveness of the Expanded Lexicon

Had we been able to use the OpinionFinder or the General Inquirer lexicons (**OFL** and **GIL**) as-is, how would the results have compared to those attained using our lexicons? We performed the baseline runs with both lexicons; OFL accuracies were 0.544-0.594 across machine learning systems, GIL's – 0.491-0.584 (see GIL column in Table 5).

We also experimented with using the *weaksubj* and *strongsubj* labels in OFL as somewhat parallel distinctions to the ones presented here (see section 4 – Related Work – for a more detailed discussion). We used (1,0,0) profile for strong positives, (0.3,0,0.7) for weak positives, (0,1,0) for strong negatives, and (0,0.3,0.7) for weak negatives, and ran all the feature representations discussed in section 5.2. Table 5 column OFL shows the best run for every machine learning system, across the different feature representations, and choosing the better performing run between vanilla OFL and the version enriched with weak/strong distinctions.

Machine Learner	**Seed BL**	**OFL**	**GIL**	**SWN**	**Exp.**
c5.0	0.512	0.598	0.491	0.516	0.641
SVM-RBF	0.527	0.594	0.495	0.520	0.644
SVM-lin.	0.584	0.594	0.580	0.569	0.630
Log. Reg.	0.545	0.598	0.541	0.537	0.616
Naïve B.	0.598	0.573	0.584	0.587	0.626

Table 5: Performance of different lexicons on essay data using various machine learning systems. For each system and lexicon, the best performance across the applicable feature representations from section 5.2 and the variants (see text) is shown. **Seed BL** column shows the best baseline performance of our seed lexicon – before paraphraser expansion and profile enrichment were applied. **Exp.** column shows the performance of Int-bin feature representation for the expanded lexicon after profile enrichment.

Additionally, we experimented with SentiWordNet (Baccianella et al., 2010). SentiWordNet is a resource for opinion mining built on top of WordNet, which assigns each synset in WordNet a score triplet (positive, negative, and objective), indicating the strength of each of these three properties for the words in the synset. The SentiWordNet annotations were automatically generated, starting with a set of manually labeled synsets. Currently, SentiWordNet includes an automatic annotation for all the synsets in WordNet, totaling more than 100,000 words. It is therefore the largest-scale lexicon with intensity information that is currently available.

Since SentiWordNet assigns scores to synsets and since our data is not sense-tagged, we induced SentiWordNet scores in the following ways. We part-of-speech tagged our train and test data using Stanford tagger (Toutanova et al., 2003). Then, we took the SentiWordNet scores for the top sense for the given part-of-speech (SWN-1). In a different variant, we took a weighted average of the scores for the different senses, using the weighting algorithm provided on SentiWordNet website[6] (SWN-2). Table 5 column SWN shows the best performance figures between SWN-1 and SWN-2, across the feature representations in section 5.2.

The comparative results in Table 5 clearly show that while our vanilla seed lexicon performs comparably to off-the-shelf lexicons on our data, the paraphraser-expanded lexicon with sentiment profiles outperforms OpinionFinder, General Inquirer, and SentiWordNet.

7.2 Sentiment Profile Enrichment with Other Lexical Expansion Methods

We presented a novel lexicon expansion method using a paraphrasing system. We also experimented with more standard methods, using WordNet and distributional similarity (Beigman Klebanov et al., 2012; Esuli and Sebastiani, 2006; Kim and Hovy, 2004; Andreevskaia and Bergler, 2006; Hu and Liu, 2004; Kanayama and Nasukawa, 2006; Strapparava and Valitutti, 2004; Kamps et al., 2004; Takamura et al., 2005; Turney and Littman, 2003; Hatzivassiloglou and McKeown, 1997). Specifically, we implemented a WordNet (Miller, 1995) based expansion that uses the 3 most frequent synonyms of the top sense of the seed word (**WN-e**). We also implemented a method based on distributional similarity: Using Lin's proximity-based thesaurus (Lin, 1998) trained on our in-house essay data as well as on well-formed newswire texts, we took all words with the proximity score $>$ 1.80 to any of the seed lexicon words (**Lin-e**). Just like the paraphraser lexicon, both perform worse than the seed lexicon in 9 out of 10 baseline runs (BL-sum and Bl-full conditions for the 5 machine learners).

To test the effect of profile enrichment, all words in WN-e and Lin-e underwent profile estimation as described in section 3.1, yielding lexicons **WN-e-p** and **Lin-e-p**, respectively. Figure 2 shows the distributions. WN-e-p and Lin-e-p exhibit similar trends to those of the paraphraser. Substituting WN-e-p for Expanded data in Table 4, we find the same relationships between the different feature sets: Int-bin>Int-sum>Int-full=BL. For Lin-e-p, Int-sum deteriorates: Int-bin>Int-sum=Int-full=BL. For the 20 runs in the Int condition, Paraphraser>WN-e-p>Lin-e-p.[7] Note that this is also the order of lexicon sizes: Lin-e is the most conservative expansion (1,907 words), WN-e is the second with 2,527 words, and the lexicon expanded using paraphrasing is the largest with 2,994 words. Table 6 shows the performance of Lin-e-p, WN-e-p, and of the Expanded lexicon from Table 4 using the Int-bin feature representation. The average relative improvements over the best baseline range between 6.6% to 14.6% for the different expansion methods.

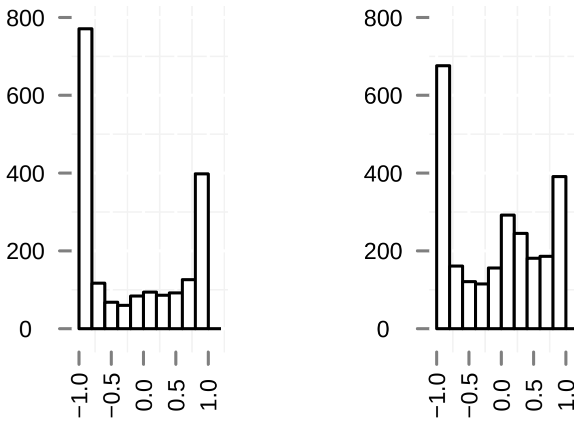

Figure 2: Sentiment profile distributions for Lin-e-p (left) and WN-e-p (right) lexicons.

Profile induction appears to be a powerful lexicon clean-up procedure that works especially well with more aggressive and thus potentially noisier expansions: The machine learners depress low-intensity and ambiguous expansions, thereby allowing the effective utilization of the improved coverage of sentiment-bearing vocabulary.

7.3 Effectiveness of the Paraphrase Expansion with Profile Enrichment Paradigm in a Different Domain

In order to check whether the paraphrase-based expasion and profile enrichment paradigm discussed in this paper generalizes to other subjectivity lexicons

[6] http://sentiwordnet.isti.cnr.it/, under "Sample code."

[7] All $>$ are signficant at p=0.05 using Wilcoxon test.

Machine Learner	Seed BL	Lin-e-p	WN-e-p	Exp.
c5.0	0.512	0.584	0.616	0.641
SVM-RBF	0.527	0.598	0.601	0.644
SVM-lin.	0.584	0.577	0.569	0.630
Log. Reg.	0.545	0.587	0.580	0.616
Naïve B.	0.598	0.591	0.623	0.626
Av. Gain		0.066	0.085	0.146

Table 6: Performance of WordNet-based, Lin-based, and Paraphraser-based expansions with profile enrichment in the Int-bin condition. **Seed BL** column shows the best baseline performance of the seed lexicon – before expansion and profile enrichment were applied. The last line shows the average relative gain over the best baseline calculated as $AG_{lex} = \Sigma_{m \in M} \frac{Lex_m - SeedBL_m}{SeedBL_m}$, where $M = \{$c5.0, SVM-RBF, SVM-linear, Logistic Regression, Naïve Bayes$\}$, and $lex \in \{$Lin-e-p, Wn-e-p, Exp$\}$.

and domains of application, we experimented with a product reviews dataset (Hu and Liu, 2004) and additional lexicons as follows.

7.3.1 Lexicons

We use the OpinionFinder and General Inquirer lexicons (OFL and GIL) as before, as well as the lexicon of positive and negative sentiment and opinion words available along with (Hu and Liu, 2004) product reviews dataset – **HL**.[8]

Since each of these lexicons contains more than 3,000 words, enrichment of the full lexicons with profiles is beyond the financial scope of our project. We therefore restrict each of the lexicons to the size of their overlap with our seed lexicon (see 2.1); the overlaps have between 415 and 467 words. These restricted lexicons are our initial lexicons for the new experiment that parallel the role of the seed lexicon in the experiments on essay data.

For each of the 3 initial lexicons **L**, L$\in\{$OFL, GIL, HL$\}$, we follow the paraphrase-based expansion as described in section 2.2. This results in about 4.5-fold expansion of each lexicon, the new lexicons **L-e**, L$\in\{$OFL, GIL, HL$\}$, numbering between 2,015 and 2,167 words. Both the initial and the expanded lexicons now undergo profile enrichment as described in section 3.1, producing lexicons **L-p** and **L-e-p**, L$\in\{$OFL, GIL, HL$\}$.

7.3.2 Data

We use the dataset from Hu and Liu (2004)[9] that contains reviews of 5 products from amazon.com: two digital cameras, a DVD player, an MP3 player, and a cellular phone. The reviews are annotated at sentence level with a label that desrcibes the particular feature that is the subject of the positive or negative evaluation and the polarity and extent of the evaluation. For example, the sentence "The phone book is very user-friendly and the speakerphone is excellent" is labeled as PHONE BOOK$[+2]$, SPEAKERPHONE$[+2]$, while the sentence "I am bored with the silver look" is labeled LOOK$[-1]$. We used all sentences that were labeled with a numerical score for at least one feature, removing a small number of sentences labeled with both positive and negative scores for different features.[10] We used the sign of the numerical score to label the sentences as positive or negative. The resulting dataset consists of 1,695 sentences, 1,061 positive and 634 negative; accuracy for a majority baseline on this dataset is 0.626. Our experiments on this dataset are done using 5-fold cross-validation.

7.3.3 Results

Table 7 shows classification accuracies for the product review data using different lexicons and machine learners. We observe that the combination of paraphrase-based expansion and profile enrichment (L-e-p column in the table) resulted in an improved performance over the initial lexicon (L column in the table) *in all cases*, with the average gain of 5% in accuracy.

Furthermore, the contributions of the expansion and the profile enrichment are complementary, since their combination performs better than each in isolation. We note that profile enrichment alone for the initial lexicon did not yield an improvement. This can be explained by the fact that the initial lexicons are highly polar, so profiles provide little additional information: The percentage of words with $\hat{p}^{pos} \geq 0.8$ or $\hat{p}^{neg} \geq 0.8$ is 84%, 86% and 91% for GIL,

[8] http://www.cs.uic.edu/∼liub/FBS/sentiment-analysis.html#lexicon

[9] http://www.cs.uic.edu/∼liub/FBS/sentiment-analysis.html#datasets, the link under "Customer Review Datasets (5 products)"

[10] such as "The headset that comes with the phone has good sound volume but it hurts the ears like you cannot imagine!"

Machine	**Lexicon Variant**			
Learner	L	L-p	L-e	L-e-p
L = OFL∩Seed, |L|=467, |L-e|=2,167				
c5.0	0.663	0.670	0.691	0.704
SVM-RBF	0.668	0.676	0.693	0.714
SVM-lin.	0.675	0.670	0.688	0.696
Log. Reg.	0.666	0.658	0.693	0.698
Naïve B.	0.668	0.668	0.686	0.695
L = GIL∩Seed, |L|=415,|L-e|=2,015				
c5.0	0.644	0.658	0.663	0.686
SVM-RBF	0.650	0.665	0.653	0.683
SVM-lin.	0.665	0.665	0.677	0.681
Log. Reg.	0.664	0.658	0.678	0.694
Naïve B.	0.669	0.666	0.678	0.703
L = HL∩Seed, |L|=434, |L-e|=2,054				
c5.0	0.676	0.675	0.689	0.706
SVM-RBF	0.673	0.674	0.700	0.713
SVM-lin.	0.676	0.664	0.703	0.710
Log. Reg.	0.668	0.661	0.703	0.699
Naïve B.	0.668	0.672	0.697	0.697

Table 7: Accuracies on product review data. For each machine learner and lexicon, the best baseline performance is shown as **L** for the initial lexicon and as **L-e** for the paraphrase-expanded lexicon. **L-p** and **L-e-p** show the performance of Int-bin feature set on the profile-enriched initial and paraphrase-expanded lexicons, respectively. The three initial lexicons L are OpinionFinder (OFL), General Inquirer (GIL), and (Hu and Liu, 2004) (HL), each intersected with our seed lexicon. Sizes of the intial and expanded lexicons are provided.

OFL, and HL-derived lexicons, respectively. In contrast, for the expanded lexicons, these percentages are 51%, 53%, and 56%; these lexicons benefit from profile enrichment.

8 Conclusions

We demonstrated a method of improving a seed sentiment lexicon by using a pivot-based paraphrasing system for lexical expansion and sentiment profile enrichment using crowdsourcing. Profile enrichment alone yielded up to 15% improvement in the performance of the seed lexicon on the task of 3-way sentence-level sentiment polarity classification of test-taker essay data. While the lexical expansion on its own failed to improve upon the performance of the seed lexicon, it became much more effective on top of sentiment profiles, generating a 7% performance boost over the best profile-enriched run with the seed lexicon. Overall, paraphrase-based expansion coupled with profile enrichment yields an up to 25% improvement in accuracy.

Additionally, we showed that our paraphrase-expanded and profile-enriched lexicon performs significantly better on our data than off-the-shelf subjectivity lexicons, namely, Opinion Finder, General Inquirer, and SentiWordNet. Furthermore, our results suggest that paraphrase-based expansion derives more benefit from profiles than two competing expansion mechanisms based on WordNet and on Lin's distributional thesaurus.

Finally, we demonstrated the effectiveness of the paraphraser-based expansion with profile enrichment paradigm on a different dataset. We used Hu and Liu (2004) product review data with sentence-level sentiment polarity labels. Paraphrase-based expansion with profile enrichment yielded an improved performance across all lexicons and machine learning algorithms we tried, with an average improvement rate of 5% in classification accuracy.

Recent literature argues that sentiment polarity is a property of word senses, rather than of words (Gyamfi et al., 2009; Su and Markert, 2008; Wiebe and Mihalcea, 2006), although Dragut et al (2012) successfully operate with "mostly negative" and "mostly positive" words based on the polarity distributions of word senses. We plan to address in future work sense disambiguation for words that have multiple senses with very different sentiment, such as *stress*, as either *anxiety* (negative) or *emphasis* (neutral).

References

Alina Andreevskaia and Sabine Bergler. 2006. Mining WordNet for a fuzzy sentiment: Sentiment tag extraction of WordNet glosses. In *Proceedings of EACL*, pages 209–216, Trento, Italy.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of LREC*, pages 2200–2204, Malta.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*, pages 597–604, Ann Arbor, MI.

Beata Beigman Klebanov, Jill Burstein, Nitin Madnani, Adam Faulkner, and Joel Tetreault. 2012. Building sentiment lexicon(s) from scratch for essay data. In *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, New Delhi, India, March.

S. Cerini, V. Compagnoni, A. Demontis, M. Formentelli, and G. Gandini. 2007. Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In Andrea Sanso, editor, *Language resources and linguistic theory: Typology, second language acquisition*, pages 200–210. Franco Angeli Editore, Milano, IT.

Fermín L. Cruz, José A. Troyano, F. Javier Ortega, and Fernando Enríquez. 2011. Automatic expansion of feature-level opinion lexicons. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 125–131, Portland, Oregon, June.

Eduard Dragut, Hong Wang, Clement Yu, Prasad Sistla, and Weiyi Meng. 2012. Polarity consistency checking for sentiment dictionaries. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 997–1005, Jeju Island, Korea, July. Association for Computational Linguistics.

Andrea Esuli and Fabrizio Sebastiani. 2006. Determining term subjectivity and term orientation for opinion mining. In *Proceedings of EACL*, pages 193–200, Trento, Italy.

Leo A. Goodman. 1965. On Simultaneous Confidence Intervals for Multinomial Proportions. *Technometrics*, 7(2):247–254.

Yaw Gyamfi, Janyce Wiebe, Rada Mihalcea, and Cem Akkaya. 2009. Integrating knowledge for subjectivity sense labeling. In *Proceedings of NAACL*, pages 10–18, Boulder, CO.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11.

Vasileios Hatzivassiloglou and Kathleen McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of ACL*, pages 174–181, Madrid, Spain.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, WA.

Valentin Jijkoun and Katja Hofmann. 2009. Generating a Non-English Subjectivity Lexicon: Relations That Matter. In *Proceedings of EACL*, pages 398–405, Athens, Greece.

Jaap Kamps, Maarten Marx, Robert Mokken, and Maarten de Rijke. 2004. Using WordNet to measure semantic orientation of adjectives. In *Proceedings of LREC*, pages 1115–1118, Lisbon, Portugal.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. In *Proceedings of EMNLP*, pages 355–363, Syndey, Australia.

Soo-Min Kim and Edward Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of COLING*, pages 1367–1373, Geneva, Switzerland.

Philip Koehn. 2005. EUROPARL: A Parallel corpus for Statistical Machine Translation. In *Proceedings of the Machine Translation Summit*.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of ACL*, pages 768–774, Montreal, Canada.

Tim Loughran and Bill McDonald. 2011. When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, 66:35–65.

Nitin Madnani and Bonnie Dorr. 2013. Generating Targeted Paraphrases for Improved Translation. *ACM Transactions on Intelligent Systems and Technology, to appear*.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of ACL*, pages 976–983, Prague, Czech Republic.

George Miller. 1995. WordNet: A lexical database. *Communications of the ACM*, 38:39–41.

Saif Mohammad, Cody Dunne, and Bonnie Dorr. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of EMNLP*, pages 599–608, Singapore, August.

Guillaume Pitel and Gregory Grefenstette. 2008. Semi-automatic building method for a multidimensional affect dictionary for a new language. In *Proceedings of LREC*, Marrakech, Morocco.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2009. Expanding domain sentiment lexicon through double propagation. In *Proceedings of the 21st international jont conference on Artifical intelligence*, IJCAI'09, pages 1199–1204.

C. Quesenberry and D. Hurst. 1964. Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics*, 6:191–195.

J. R. Quinlan. 1993. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers.

Delip Rao and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of EACL*, pages 675–682, Athens.

Philip Stone, Dexter Dunphy, Marshall Smith, and Daniel Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-affect: an affective extension of WordNet. In *Proceedings of LREC*, pages 1083–1086, Lisbon, Portugal.

Fangzhong Su and Katja Markert. 2008. Eliciting Subjectivity and Polarity Judgements on Word Senses. In *Proceedings of COLING*, pages 825–832, Manchester, UK.

P. Subasic and A. Huettner. 2001. Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems*, 9(4).

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-Based Method for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientation of words using spin model. In *Proceedings of ACL*, pages 133–140, Ann Arbor, MI.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL*, pages 252–259.

Peter Turney and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315346.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of Web-derived polarity lexicons. In *Proceedings of NAACL*, pages 777–785, Los Angeles, CA.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of ACL*, pages 1065–1072, Sydney, Australia.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of CICLING (invited paper)*, pages 486–497, Mexico City.