Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus

Saif Mohammad[φ†*]**, Cody Dunne**[◇‡]**,** and **Bonnie Dorr**[φ†‡*]
Laboratory for Computational Linguistics and Information Processing[φ]
Human-Computer Interaction Lab[◇]
Institute for Advanced Computer Studies[†]
Department of Computer Science[‡], University of Maryland.
Human Language Technology Center of Excellence[*]
{saif,bonnie}@umiacs.umd.edu and {cdunne}@cs.umd.edu

Abstract

Sentiment analysis often relies on a semantic orientation lexicon of positive and negative words. A number of approaches have been proposed for creating such lexicons, but they tend to be computationally expensive, and usually rely on significant manual annotation and large corpora. Most of these methods use WordNet. In contrast, we propose a simple approach to generate a high-coverage semantic orientation lexicon, which includes both individual words and multi-word expressions, using only a Roget-like thesaurus and a handful of affixes. Further, the lexicon has properties that support the Polyanna Hypothesis. Using the General Inquirer as gold standard, we show that our lexicon has 14 percentage points more correct entries than the leading WordNet-based high-coverage lexicon (SentiWordNet). In an extrinsic evaluation, we obtain significantly higher performance in determining phrase polarity using our thesaurus-based lexicon than with any other. Additionally, we explore the use of visualization techniques to gain insight into the our algorithm beyond the evaluations mentioned above.

1 Introduction

Sentiment analysis involves determining the opinions and private states (beliefs, emotions, speculations, and so on) of the speaker (Wiebe, 1994). It has received significant attention in recent years due to increasing online opinion content and applications in tasks such as automatic product recommendation systems (Tatemura, 2000; Terveen et al., 1997), question answering (Somasundaran et al., 2007; Lita et al., 2005), and summarizing multiple view points (Seki et al., 2004) and opinions (Mohammad et al., 2008a).

A crucial sub-problem is to determine whether positive or negative sentiment is expressed. Automatic methods for this often make use of lexicons of words tagged with positive and negative semantic orientation (Turney, 2002; Wilson et al., 2005; Pang and Lee, 2008). A word is said to have a positive **semantic orientation (SO)** (or **polarity**) if it is often used to convey favorable sentiment or evaluation of the topic under discussion. Some example words that have positive semantic orientation are *excellent, happy, honest*, and so on. Similarly, a word is said to have negative semantic orientation if it is often used to convey unfavorable sentiment or evaluation of the target. Examples include *poor, sad,* and *dishonest*.

Certain semantic orientation lexicons have been manually compiled for English—the most notable being the **General Inquirer (GI)** (Stone et al., 1966).[1] However, the GI lexicon has orientation labels for only about 3,600 entries. The **Pittsburgh subjectivity lexicon (PSL)** (Wilson et al., 2005), which draws from the General Inquirer and other sources, also has semantic orientation labels, but only for about 8,000 words.

Automatic approaches to creating a semantic orientation lexicon and, more generally, approaches for word-level sentiment annotation can be grouped into two kinds: (1) those that rely on manually created lexical resources—most of which use WordNet (Strapparava and Valitutti, 2004; Hu and Liu, 2004; Kamps et al., 2004; Takamura et al., 2005; Esuli and Sebastiani, 2006; An-

---

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

dreevskaia and Bergler, 2006; Kanayama and Nasukawa, 2006); and (2) those that rely on text corpora (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Yu and Hatzivassiloglou, 2003; Grefenstette et al., 2004). Many of these lexicons, such as **SentiWordNet (SWN)** (Esuli and Sebastiani, 2006) were created using supervised classifiers and significant manual annotation, whereas others such as the **Turney and Littman lexicon (TLL)** (2003) were created from very large corpora (more than 100 billion words).

In contrast, we propose a computationally inexpensive method to compile a high-coverage semantic orientation lexicon without the use of any text corpora or manually annotated semantic orientation labels. Both of these resources may be used, if available, to further improve results. The lexicon has about twenty times the number of entries in the GI lexicon, and it includes entries for both individual words and common multi-word expressions. The method makes use of a Roget-like thesaurus and a handful of antonym-generating affix patterns. Whereas thesauri have long been used to estimate semantic distance (Jarmasz and Szpakowicz, 2003; Mohammad and Hirst, 2006), the closest thesaurus-based work on sentiment analysis is by Aman and Szpakowicz (2007) on detecting emotions such as happiness, sadness, and anger. We evaluated our thesaurus-based algorithm both intrinsically and extrinsically and show that the semantic orientation lexicon it generates has significantly more correct entries than the state-of-the-art high-coverage lexicon SentiWordNet, and that it has a significantly higher coverage than the General Inquirer and Turney–Littman lexicons.

In Section 2 we examine related work. Section 3 presents our algorithm for creating semantic orientation lexicons. We describe intrinsic and extrinsic evaluation experiments in Section 4, followed by a discussion of the results in Section 5. Additionally, in Section 6 we show preliminary visualizations of how our algorithm forms chains of positive and negative thesaurus categories. Good visualizations are not only effective in presenting information to the user, but also help us better understand our algorithm. Section 7 has our conclusions.

2 Related Work

Pang and Lee (2008) provide an excellent survey of the literature on sentiment analysis. Here we briefly describe the work closest to ours.

Hatzivassiloglou and McKeown (1997) proposed a supervised algorithm to determine the semantic orientation of adjectives. They first generate a graph that has adjectives as nodes. An edge between two nodes indicates either that the two adjectives have the same or opposite semantic orientation. A clustering algorithm partitions the graph into two subgraphs such that the nodes in a subgraph have the same semantic orientation. The subgraph with adjectives that occur more often in text is marked positive and the other negative. They used a 21 million word corpus and evaluated their algorithm on a labeled set of 1336 adjectives (657 positive and 679 negative). Our approach does not require manually annotated semantic orientation entries to train on and is much simpler.

Esuli and Sebastiani (2006) used a supervised algorithm to attach semantic orientation scores to WordNet glosses. They train a set of ternary classifiers using different training data and learning methods. The set of semantic orientation scores of all WordNet synsets is released by the name SentiWordNet.[2] An evaluation of SentiWordNet by comparing orientation scores of about 1,000 WordNet glosses to scores assigned by human annotators is presented in Esuli (2008). Our approach uses a Roget-like thesaurus, and it does not use any supervised classifiers.

Turney and Littman (2003) proposed a minimally supervised algorithm to calculate the semantic orientation of a word by determining if its tendency to co-occur with a small set of positive words is greater than its tendency to co-occur with a small set of negative words. They show that their approach performs better when it has a large amount of text at its disposal. They use text from 350 million web-pages (more than 100 billion words). Our approach does not make use of any text corpora, although co-occurrence statistics could be used to further improve the lexicon. Furthermore, our lexicon has entries for commonly used multi-word expressions as well.

Mohammad et al. (2008b) developed a method to determine the degree of antonymy (contrast) between two words using the *Macquarie The-*

---

saurus (Bernard, 1986), co-occurrence statistics, and a small set of antonym-generating affix patterns such as X–*dis*X. Often, one member of a pair of contrasting terms is positive and one member is negative. In this paper, we describe how a subset of those affix patterns can be used in combination with a thesaurus and the edicts of marking theory to create a large lexicon of words and phrases marked with their semantic orientation.

3 Generating the Semantic Orientation Lexicon

Our algorithm to generate a semantic orientation lexicon has two steps: (1) identify a seed set of positive and negative words; (2) use a Roget-like thesaurus to mark the words synonymous with the positive seeds "positive" and words synonymous with the negative seeds "negative". The two steps are described in the subsections below. Our implementation of the algorithm used the *Macquarie Thesaurus* (Bernard, 1986). It has about 100,000 unique words and phrases.

3.1 Seed words

3.1.1 Automatically identifying seed words

It is known from marking theory that overtly marked words, such as *dishonest, unhappy,* and *impure*, tend to have negative semantic orientation, whereas their unmarked counterparts, *honest, happy,* and *pure*, tend to have positive semantic orientation (Lehrer, 1974; Battistella, 1990). Exceptions such as *biased–unbiased* and *partial–impartial* do exist, and in some contexts even a predominantly negative marked word may be positive. For example *irreverent* is negative in most contexts, but positive in the sentence below:

> *Millions of fans follow Moulder's irreverent quest for truth.*

However, as we will show through experiments, the exceptions are far outnumbered by those that abide by the predictions of marking theory.

We used a set of 11 antonym-generating affix patterns to generate overtly marked words and their unmarked counterparts (Table 1). Similar antonyms-generating affix patterns exist for many languages (Lyons, 1977). The 11 chosen affix patterns generated 2,692 pairs of marked and unmarked valid English words that are listed in the *Macquarie Thesaurus*. The marked words

Affix pattern		# word	
w_1	w_2	pairs	example word pair
X	*dis*X	382	*honest–dishonest*
X	*im*X	196	*possible–impossible*
X	*in*X	691	*consistent–inconsistent*
X	*mal*X	28	*adroit–maladroit*
X	*mis*X	146	*fortune–misfortune*
X	*non*X	73	*sense–nonsense*
X	*un*X	844	*happy–unhappy*
X	X*less*	208	*gut–gutless*
*l*X	*ill*X	25	*legal–illegal*
*r*X	*ir*X	48	*responsible–irresponsible*
X*less*	X*ful*	51	*harmless–harmful*
Total		2692	

Table 1: Eleven affix patterns used to generate the seed set of marked and unmarked words. Here 'X' stands for any sequence of letters common to both words w_1 and w_2.

are deemed negative and the unmarked ones positive, and these form our seed set of positive and negative words. We will refer to this set of orientation-marked words as the **affix seeds lexicon (ASL)**. Note that some words may have multiple marked counterparts, for example, *trust–trustless* and *trust–mistrust*. Thus, ASL has more negative words (2,652) than positive ones (2,379). Also, the X*less*–X*ful* pattern generates word pairs that are both overtly marked; words generated from X*less* are deemed negative and words generated from X*ful* are deemed positive.

It should be noted that the affix patterns used here are a subset of those used in Mohammad et al. (2008b) to generate antonym pairs. The affix patterns ignored are those that are not expected to generate pairs of words with opposite semantic orientation. For instance, the pattern *imX-exX* generates word pairs such as *import–export* and *implicit–explicit* that are antonymous, but do not have opposite semantic orientations.

3.1.2 Using manually annotated seed words

Since manual semantic orientation labels exist for some English words (the GI lexicon), we investigated their usefulness in further improving the coverage and correctness of the entries in our lexicon. We used the GI words as seeds in the same way as the words generated from the affix patterns were used (Section 3.1.1).

3.2 Generalizing from the seeds

A published thesaurus such as the *Roget's* or *Macquarie* has about 1,000 categories, each consisting of on average 120 words and commonly used

SO lexicon	Mode of creation	Resources used	# entries	# positives	# negatives
ASL	automatic	11 affix rules	5,031	2,379 (47.3%)	2,652 (52.7%)
GI	manual	human SO annotation	3,605	1,617 (44.9%)	1,988 (55.1%)
GI-subset	manual	human SO annotation	2,761	1,262 (45.7%)	1,499 (54.3%)
MSOL(ASL)	automatic	thesaurus, 11 affix rules	51,157	34,152 (66.8%)	17,005 (33.2%)
MSOL(GI)	automatic	GI, thesaurus	69,971	25,995 (37.2%)	43,976 (62.8%)
MSOL(ASL and GI)	automatic	GI, thesaurus, 11 affix rules	76,400	30,458 (39.9%)	45,942 (60.1%)
PSL	mostly manual	GI, other sources	6,450	2,298 (35.6%)	4,485 (64.4%)
SWN	automatic	human SO annotation, WordNet, ternary classifiers	56,200	47,806 (85.1%)	8,394 (14.9%)
TLL	automatic	100 billion word corpus, minimal human SO annotation	3,596	1,625 (45.2%)	1,971 (54.8%)

Table 2: Key details of semantic orientation (SO) lexicons. ASL = affix seeds lexicon, GI = General Inquirer, MSOL = Macquarie semantic orientation lexicon, PSL = Pitt subjectivity lexicon, SWN = SentiWordNet, TLL = Turney–Littman lexicon.

multi-word expressions. Terms within a category tend to be closely related, and they are further grouped into sets of near-synonymous words and phrases called **paragraphs**. There are about 10,000 paragraphs in most Roget-like thesauri.

Every thesaurus paragraph is examined to determine if it has a seed word (by looking up the seed lexicon described in Section 3.1). If a thesaurus paragraph has more positive seed words than negative seed words, then all the words (and multi-word expressions) in that paragraph are marked as positive. Otherwise, all its words are marked negative.

Note that this method assigns semantic orientation labels to word–**thesaurus paragraph** pairs. Thesaurus paragraphs can be thought of as word senses. A word with multiple meanings is listed in multiple thesaurus paragraphs, and so will be assigned semantic orientation labels for each of these paragraphs. Thus, the method assigns a semantic orientation to a word–sense combination similar to the SentiWordNet approach and differing from the General Inquirer and Turney–Littman lexicons.

However, in most natural language tasks, the intended sense of the target word is not explicitly marked. So we generate a word-based lexicon by asking the different senses of a word to vote. If a word is listed in multiple thesaurus paragraphs, then the semantic orientation label most common to them is chosen as the word's label. We will refer to this set of word–**semantic orientation** pairs as the **Macquarie Semantic Orientation Lexicon (MSOL)**. A set created from only the affix seeds will be called **MSOL(ASL)**, a set created from only the GI seeds will be called **MSOL(GI)**, and

the set created using both affix seeds and GI seeds will be called **MSOL(ASL and GI)**.[3] We generated a similar word-based lexicon for SentiWordNet (SWN) by choosing the semantic orientation label most common to the synsets pertaining to a target word.

Table 2 summarizes the details of all the lexicons. MSOL(ASL and GI) has a much larger percentage of negatives than MSOL(ASL) because GI has a much larger percentage of negative words. These negative seeds generate many more negative entries in MSOL(ASL and GI). Of the 51,157 entries in MSOL(ASL), 47,514 are single-word entries and 3,643 are entries for multi-word expressions. Of the 69,971 entries in MSOL(GI), 45,197 are single-word entries and 24,774 are entries for common multi-word expressions. Of the 76,400 entries in MSOL(ASL and GI), 51,208 are single-word entries and 25,192 are entries for common multi-word expressions. In our evaluation, we used only the single-word entries to maintain a level playing field with other lexicons.

4 Evaluation

We evaluated the semantic orientation lexicons both intrinsically (by comparing their entries to the General Inquirer) and extrinsically (by using them in a phrase polarity annotation task).

4.1 Intrinsic: Comparison with GI

Similar to how Turney and Littman (2003) evaluated their lexicon (TLL), we determine if the semantic orientation labels in the automatically generated lexicons match the semantic orientation la-

---

[3]MSOL is publicly available at: www.umiacs.umd.edu/ ∼saif/WebPages/ResearchInterests.html.

Lexicon	All	Positives	Negatives
MSOL(ASL)	74.3	84.2	65.9
SWN	60.1	86.5	37.9
TLL	**83.3**	83.8	82.8

Table 3: The percentage of GI-subset entries (all, only the positives, only the negatives) that match those of the automatically generated lexicons.

bels of words in GI. GI, MSOL(ASL), SWN, and TLL all have 2,761 words in common. We will call the corresponding 2,761 GI entries the **GI-subset**.

Table 3 shows the percentage of GI-subset entries that match those of the three automatically-generated lexicons (MSOL(ASL), SWN, and TLL). (The differences in percentages shown in the table are all statistically significant—$p <$ 0.001.) We do not show results for MSOL(GI), MSOL(ASL and GI), and the Pittsburgh subjectivity lexicon (PSL) because these lexicons were created using GI entries. TLL most closely matches the GI-subset, and MSOL matches the GI-subset more closely than SWN with the GI-subset. However, the goal of this work is to produce a high-coverage semantic orientation lexicon and so we additionally evaluate the lexicons on the extrinsic task described below.

4.2 Extrinsic: Identifying phrase polarity

The MPQA corpus contains news articles manually annotated for opinions and private states.[4] Notably, it also has polarity annotations (positive/negative) at the phrase-level. We conducted an extrinsic evaluation of the manually-generated and automatically-generated lexicons by using them to determine the polarity of phrases in the MPQA version 1.1 collection of positive and negative phrases (1,726 positive and 4,485 negative).

We used a simple algorithm to determine the polarity of a phrase: (1) If any of the words in the target phrase is listed in the lexicon as having negative semantic orientation, then the phrase is marked negative. (2) If none of the words in the phrase is negative and if there is at least one positive word in the phrase, then it is marked positive. (3) In all other instances, the classifier refrains from assigning a tag. Indeed better accuracies in phrase semantic orientation annotation can be obtained by using supervised classifiers and more sophisticated context features (Choi and Cardie,

---
[4]http://www.cs.pitt.edu/mpqa

2008). However, our goal here is only to use this task as a testbed for evaluating different semantic orientation lexicons, and so we use the method described above to avoid other factors from influencing the results.

Table 4 shows the performance of the algorithm when using different lexicons. The performance when using lexicons that additionally make use of GI entries—MSOL(GI), MSOL(ASL and GI), PSL, and a combined GI-SWN lexicon—is shown lower down in the table. GI–SWN has entries from both GI and SWN. (For entries with opposing labels, the GI label was chosen since GI entries were created manually.) Observe that the best F-scores are obtained when using MSOL (in both categories—individual lexicons and combinations with GI). The values are significantly better than those attained by others ($p < 0.001$).

5 Discussion

The extrinsic evaluation shows that our thesaurus- and affix-based lexicon is significantly more accurate than SentiWordNet. Moreover, it has a much larger coverage than the GI and Pitt lexicons. Observe also that the affix seeds set, by itself, attains only a modest precision and a low recall. This is expected because it is generated by largely automatic means. However, the significantly higher MSOL performance suggests that the generalization step (described in Section 3.2) helps improve both precision and recall. Precision is improved because multiple seed words vote to decide the semantic orientation of a thesaurus paragraph. Recall improves simply because non-seed words in a paragraph are assigned the semantic orientation that is most prevalent among the seeds in the paragraph.

5.1 Support for the Polyanna Hypothesis

Boucher and Osgood's (1969) Polyanna Hypothesis states that people have a preference for using positive words and expressions as opposed to using negative words and expressions. Studies have shown that indeed speakers across languages use positive words much more frequently than negative words (Kelly, 2000). The distribution of positive and negative words in MSOL(ASL) further supports the Polyanna Hypothesis as it shows that even if we start with an equal number of positive and negative seed words, the expansion of the positive set through the thesaurus is much more pro-

	All phrases			Only positives			Only negatives		
SO lexicon	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Individual lexicons									
ASL	0.451	0.165	0.242	0.451	0.165	0.242	0.192	0.063	0.095
GI	0.797	0.323	0.459	0.871	0.417	0.564	0.763	0.288	0.419
MSOL(ASL)	0.623	0.474	**0.539**	0.631	0.525	0.573	0.623	0.458	**0.528**
SWN	0.541	0.408	0.465	0.745	0.624	**0.679**	0.452	0.328	0.380
TLL	0.769	0.298	0.430	0.761	0.352	0.482	0.775	0.279	0.411
Automatic lexicons + GI information									
MSOL(GI)	0.713	0.540	0.615	0.572	0.470	0.516	0.777	0.571	**0.658**
MSOL(ASL and GI)	0.710	0.546	**0.617**	0.577	0.481	0.525	0.771	0.574	**0.658**
PSL	0.823	0.422	0.558	0.860	0.487	0.622	0.810	0.399	0.535
GI-SWN	0.650	0.494	0.561	0.740	0.623	**0.677**	0.612	0.448	0.517

Table 4: Performance in phrase polarity tagging. P = precision, R = recall, F = balanced F-score. The best F-scores in each category are marked in bold.

nounced than the expansion of the negative set. (About 66.8% of MSOL(ASL) words are positive, whereas only 33.2% are negative.) This suggests that there are many more near-synonyms of positive words than near-synonyms of negative ones, and so there are many more forms for expressing positive sentiments than forms for expressing negative sentiment.

5.2 Limitations

Some of the errors in MSOL were due to non-antonymous instantiations of the affix patterns. For example, *immigrate* is not antonymous to *migrate*. Other errors occur because occasionally the words in the same thesaurus paragraph have differing semantic orientations. For example, one paragraph has the words *slender* and *slim* (which, many will agree, are positive) as well as the words *wiry* and *lanky* (which many will deem negative). Both these kinds of errors can be mitigated using a complementary source of information, such as co-occurrence with other known positive and negative words (the Turney–Littman method).

5.3 Future work

Theoretically, a much larger Turney–Littman lexicon can be created even though it may be computationally intensive when working with 100 billion words. However, MSOL and TLL are created from different sources of information—MSOL from overtly marked words and a thesaurus, and TLL from co-occurrence information. Therefore, a combination of the two approaches is expected to produce an even more accurate semantic orientation lexicon, even with a modest-sized corpus at its disposal. This is especially attractive for low resource languages. We are also developing meth-ods to leverage the information in an English thesaurus to create semantic orientation lexicons for a low-resource language through the use of a bilingual lexicon and a translation disambiguation algorithm.

6 Visualizing the semantic orientation of thesaurus categories

In recent years, there have been substantial developments in the field of information visualization, and it is becoming increasingly clear that good visualizations can not only convey information quickly, but are also an important tool for gaining insight into an algorithm, detecting systematic errors, and understanding the task. In this section, we present some preliminary visualizations that are helping us understand our approach beyond the evaluations described above.

As discussed in Section 3.1.1, the affix seeds set connects the thesaurus words with opposite semantic orientation. Usually these pairs of words occur in different thesaurus categories, but this is not necessary. We can think of these connections as relationships of contrast in meaning and semantic orientation, not just between the two words but also between the two categories. To better aid our understanding of the automatically determined category relationships we visualized this network using the Fruchterman-Reingold force-directed graph layout algorithm (Fruchterman and Reingold, 1991) and the NodeXL network analysis tool (Smith et al., 2009) [5].

Our dataset consists of 812 categories from the *Macquarie Thesaurus* and 27,155 antonym edges between them. There can be an edge from a cat-

---
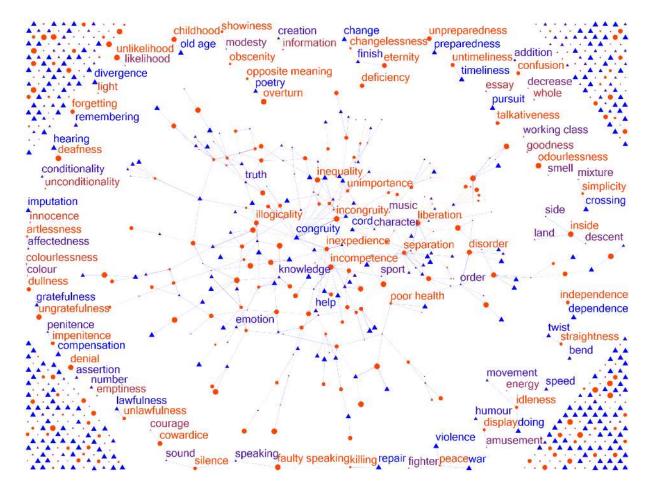[5] Available from http://www.codeplex.com/NodeXL

Figure 1: After removing edges with low weight we can see the structure the network backbone. Isolate category pairs are drawn in a ring around the main connected component and singletons are staggered in the corners. Each node is colored by its semantic orientation (red for negative, blue for positive) and edges are colored by their weight, from red to blue. Node shape also codes semantic orientation, with triangles positive and circles negative. Size codes the magnitude the semantic orientation, with the largest nodes representing the extremes. Node labels are shown for nodes in isolates and those in the top 20 for betweenness centrality.

egory to itself called a self-edge, indicating that a word and its antonym (with opposite semantic orientation) both exist in the same category. There can be multiple edges between two categories indicating that one or more words in one category have one or more antonyms in the other category. These multiple edges between category pairs were merged together resulting in 14,597 weighted meta-edges. For example, if there are n edges between a category pair they were replaced by a single meta-edge of weight n.

The network is too dense and interconnected for force-directed placement to generate a useful publication-size drawing of the entire network. By removing edges that had a weight less than 6, we can visualize a smaller and more understandable 540 edge network of the core categories and any

new isolates created. Additionally, we show only edges between categories with opposite semantic orientations (Figure 1). Observe that there are three groups of nodes: those in the core connected component, the small isolates in the ring surrounding it, and the connectionless singletons arranged in the corners.

Each node c (thesaurus category) is colored on a red to blue continuous scale according to its semantic orientation SO, which is computed purely from its graph structure (in-degree ID and out-degree OD):

$$SO(c) = \frac{OD(c) - ID(c)}{OD(c) + ID(c)} \qquad (1)$$

Blue nodes represent categories with many positive words; we will call them **positive cate-**

gories (p). Red nodes are categories with many negative words; we will call them **negative categories (n)**. Shades of purple in between are categories that have words with both positive and negative semantic orientation (**mixed categories**). Similarly, edges are colored according to their weight from red (small weight) to blue (large weight). We also use shape coding for semantic orientation, with triangles being positive and circles negative, and the size of the node depicts the magnitude of the semantic orientation. For example, the pair HEARING(p)–DEAFNESS(n) in the top left of Figure 1 represent the two size extremes: HEARING has a semantic orientation of 1 and DEAFNESS has a score of -1. The mixed categories with near 0 semantic orientation such as LIKELIHOOD with a score of .07 are the smallest.

Nodes are labeled by the thesaurus-provided head words—a word or phrase that best represents the coarse meaning of the category. For readability, we have restricted the labels to nodes in the isolates and the top 20 nodes in the core connected component that have the highest **betweenness centrality**, which means they occur on more shortest paths between other nodes in the network (i.e., they are the bridges or gatekeepers).

From the ring of isolates we can see how many antonymous categories, and their semantic orientations, are correctly recognized. For example, ASSERTION(p)–DENIAL(n), HEARING(p)–DEAFNESS(n), GRATEFULNESS(p)–UNGRATEFULNESS(n), and so on. Some codings may seem less intuitive, such as those in the core, but much of this is the effect of abstracting away the low weight edges, which may have more clearly identified the relationships.

An alternative approach to removing edges with low weight is to filter categories in the network based on graph-theoretic metrics like betweenness centrality, **closeness centrality**, and **eigenvector centrality**. We discussed betweenness centrality before. The closeness centrality of a node is the average distance along the shortest path between that node and all other nodes reachable from it. Eigenvector centrality is another measure of node importance, assigning node score based on the idea that connections to high-scoring nodes are more important than those to low-scoring ones. We removed nodes with less than 0.1 betweenness centrality, less than 0.04 eigenvector centrality, and above 2.1 closeness centrality, leaving

the key 56 nodes. They have 497 edges between them, of which we show only those between categories with opposite semantic orientations (Figure 2). Node and edge color, size, and shape coding is as before.

Observe that most of these categories have a strongly evaluative nature. Also, as our algorithm makes connections using overt negative markers, it makes sense that the central categories in our network have negative orientation (negative categories have many words with overt markings). It is interesting, though, how some positive and mixed categories reside in the core too. Further inspection revealed that these categories have a large number of words within them. For example, it may be less intuitive as to why the category of MUSIC is listed in the core, but this is because it has about 1,200 words in it (on average, each category has about 120 words), and because many of these words, such as *harmonious(p), melodious(n),* and *lament(n)* are evaluative in nature.

7 Conclusion

We created a high-coverage semantic orientation lexicon using only affix rules and a Roget-like thesaurus. The method does not require terms with manually annotated semantic orientation labels, though we show that if available they can be used to further improve both the correctness of its entries and its coverage. The lexicon has about twenty times as many entries as in the General Inquirer and the Turney–Littman lexicons, and includes entries for both individual words and common multi-word expressions. Experiments show that it has significantly more correct entries than SentiWordNet. The approach is complementary to that of Turney and Littman (2003) and a combination of this approach with co-occurrence statistics (even if drawn from a modest sized corpus) is expected to yield an even better lexicon.

Visualization of the thesaurus categories as per the semantic orientations assigned to them by our algorithm reveals that affix patterns produce a strongly connected graph and that indeed there are many long chains of positive and negative categories. Furthermore, the key categories of this graph (the ones with high centrality and closeness) are strongly evaluative in nature, and most of them tend to have negative semantic orientation.
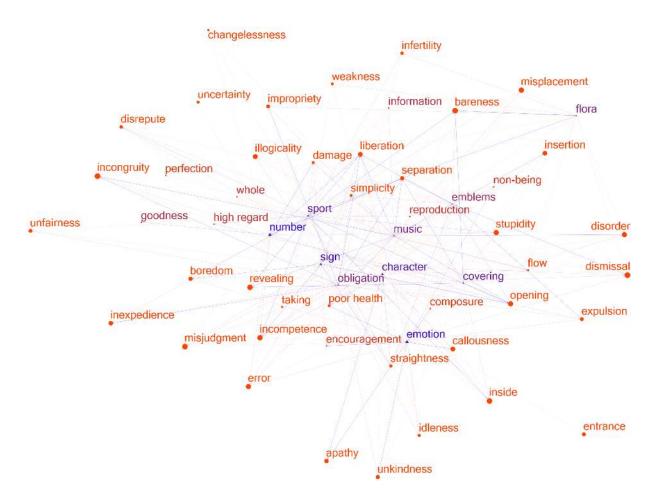
Figure 2: After filtering out nodes based on graph-theoretic metrics, the core of the network becomes visible. The visualization is colored as in Figure 1, and we can see how the core is dominated by categories with negative semantic orientation (red). Shape, size, and color coding is as before.

Acknowledgments

References

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. *Text, Speech and Dialogue*, 4629:196–205.

Alina Andreevskaia and Sabine Bergler. 2006. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings of the EACL*, Trento, Italy.

Edwin Battistella. 1990. *Markedness: The Evaluative Superstructure of Language*. State University of New York Press, Albany, New York.

John R. L. Bernard, editor. 1986. *The Macquarie Thesaurus*. Macquarie Library, Sydney, Australia.

Jerry D. Boucher and Charles E. Osgood. 1969. The pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behaviour*, 8:1–8.

Yejin Choi and Claire Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP-2008)*, Waikiki, Hawaii.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, pages 417–422, Genoa, Italy.

Andrea Esuli. 2008. *Automatic Generation of Lexical Resources for Opinion Mining: Models, Algorithms and Applications*. Ph.D. thesis, Department of Information Engineering, University of Pisa, Pisa, Italy.

Thomas M. J. Fruchterman and Edward M. Reingold. 1991. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164.

Gregory Grefenstette, Yan Qu, David Evans, and James Shanahan. 2004. Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes. In James Shanahan Yan Qu and Janyce Wiebe, editors, *Exploring Attitude and Affect in Text: Theories and Applications, AAAI-2004 Spring Symposium Series*, pages 71–78, San Jose, California.

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

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD International ConferenceDiscovery and Data Mining (KDD-04)*, Seattle, WA.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's Thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, pages 212–219, Borovets, Bulgaria.

Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. 2004. Using WordNet to measure semantic orientation of adjectives. In *LREC*.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 355–363, Sydney, Australia, July. Association for Computational Linguistics.

Michael H. Kelly. 2000. Naming on the bright side of life. volume 48, pages 3–26.

Adrienne Lehrer. 1974. *Semantic fields and lexical structure*. North-Holland; American Elsevier, Amsterdam and New York.

Lucian Vlad Lita, Andrew Hazen Schlaikjer, WeiChang Hong, and Eric Nyberg. 2005. Qualitative dimensions in question answering: Extending the definitional QA task. In *Proceedings of AAAI*, pages 1616–1617. Student abstract.

John Lyons. 1977. *Semantics*, volume 1. Cambridge University Press.

Saif Mohammad and Graeme Hirst. 2006. Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, pages 35–43, Sydney, Australia.

Saif Mohammad, Bonnie Dorr, Melissa Egan, Jimmy Lin, and David Zajic. 2008a. Multiple alternative sentence compressions and word-pair antonymy for automatic text summarization and recognizing textual entailment. In *Proceedings of the Text Analysis Conference (TAC-2008)*, Gaithersburg, MD.

Saif Mohammad, Bonnie Dorr, and Graeme Hirst. 2008b. Computing word-pair antonymy. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Waikiki, Hawaii.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Yohei Seki, Koji Eguchi, and Noriko Kando. 2004. Analysis of multi-document viewpoint summarization using multi-dimensional genres. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 142–145.

Marc Smith, Ben Shneiderman, Natasa Milic-Frayling, Eduarda Mendes Rodrigues, Vladimir Barash, Cody Dunne, Tony Capone, Adam Perer, and Eric Gleave. 2009. Analyzing (social media) networks with NodeXL. In *C&T '09: Proc. Fourth International Conference on Communities and Technologies*, Lecture Notes in Computer Science. Springer.

Swapna Somasundaran, Theresa Wilson, Janyce Wiebe, and Veselin Stoyanov. 2007. QA with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

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

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

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientation of words using spin model. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 133–140.

Junichi Tatemura. 2000. Virtual reviewers for collaborative exploration of movie reviews. In *Proceedings of Intelligent User Interfaces (IUI)*, pages 272–275.

Loren Terveen, Will Hill, Brian Amento, David McDonald, and Josh Creter. 1997. PHOAKS: A system for sharing recommendations. *Communications of the Association for Computing Machinery (CACM)*, 40(3):59–62.

Peter Turney and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL*, pages 417–424, Philadelphia, Pennsylvania.

Janyce M. Wiebe. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

Theresa Wilson, Janyce Wiebe, and Paul Hoffman. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, pages 347–354, Vancouver, Canada.

Hong Yu and Vassileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP*, pages 129–136, Morristown, NJ.