Building Subjectivity Lexicon(s) From Scratch For Essay Data

Beata Beigman Klebanov,[1] Jill Burstein,[1] Nitin Madnani,[1]
Adam Faulkner,[2] Joel Tetreault[1]

[1] Educational Testing Service
{bbeigmanklebanov,jburstein,nmadnani,jtetreault}@ets.org
[2] Graduate Center, The City University of New York
adamflkr@gmail.com

Abstract. While there are a number of subjectivity lexicons available for research purposes, none can be used commercially. We describe the process of constructing subjectivity lexicon(s) for recognizing sentiment polarity in essays written by test-takers, to be used within a commercial essay-scoring system. We discuss ways of expanding a manually-built seed lexicon using dictionary-based, distributional in-domain and out-of-domain information, as well as using Amazon Mechanical Turk to help “clean up” the expansions. We show the feasibility of constructing a family of subjectivity lexicons from scratch using a combination of methods to attain competitive performance with state-of-art research-only lexicons. Furthermore, this is the first use, to our knowledge, of a paraphrase generation system for expanding a subjectivity lexicon.

Keywords: essay writing, sentiment analysis, sentiment polarity, subjectivity lexicon, C5.0, lexicon expansion, paraphrase generation, thesaurus resources.

1 Introduction

For commercial applications of sentiment analysis, an in-house subjectivity lexicon needs to be constructed, since existing lexicons, such as MPQA [1] and GI [2], are available either for research and education only[1] or under GNU GPL license that disallows the incorporation of the resource into proprietary materials.[2] In this article, we describe a methodology for creating a family of subjectivity lexicons from scratch through the following phases: (1) a lexicon of about 400 words was manually

[1] “This version of the General Inquirer is made available exclusively for educational and research purposes.” From http://www.wjh.harvard.edu/~inquirer/j1_1/manual.

[2] “The GNU General Public License does not permit incorporating your program into proprietary programs.” From http://www.gnu.org/copyleft/gpl.html.

constructed based on materials in our domain of interest (test-taker essays), (2) a small-scale annotation was conducted to augment the lexicon to 750 words, and (3) a variety of expansion methods with subsequent *human* and *automated* clean-up were implemented. We show that this process results in subjectivity lexicons that are comparable to state-of-art lexicons in terms of sentiment classification performance on our data as well as in terms of effective coverage (the number of words in a lexicon that appear in our data).

The article is organized as follows. Section 2 details the process of lexicon construction, starting from the 750-word seed lexicon (section 2.1), then discussing the automatic lexicon expansions (section 2.2), proceeding to the manual clean-up using Amazon Mechanical Turk (**AMT**) (section 2.3) and automatic clean-up through lexicon combination (section 2.4). Section 3 details the evaluation of the lexicons; the setup for evaluation is described in sections 3.1 and 3.2, section 3.3 compares the lexicons in terms of effective coverage of our data, section 3.4 provides the comparative evaluation of the lexicons on the sentence-level sentiment classification task. Table 4 in section 3.4 presents our main results. Section 4 surveys related work. We discuss our results and conclude in section 5.

2 Building Subjectivity Lexicons

2.1 Seed Lexicon

First, we randomly sample 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 accounting exams. We manually selected positive and negative sentiment words from the full list of word types in these data; these constitute our **Lexicon 0**, which contains 407 words.

We sampled 878 sentences containing at least one word from Lexicon 0, thus biasing the sample towards sentiment-bearing sentences. The motivation for the bias was increasing the incidence of sentiment-bearing – positive (**POS**) and negative (**NEG**) – sentences, under the assumption that sentiment-bearing sentences had more positive and negative words, and hence, were more effective for lexicon development. Using these sentences, we proceeded with an annotation task as follows. Two research assistants annotated 878 sentences with sentence-level sentiment polarity; 248 of these were also annotated for all words that contribute to the sentiment of the sentence or go against it. We refer to the 248 sentence set as **L-1**, and to the 630 sentence set **L-2**. For example, the following sentence was labeled as positive; words contributing to the positive sentiment are bold-faced and words (and phrases) going against it are underlined.

Some may even be **impressed** that we are **confident** enough to risk showing a lower net income.

In addition, positive and negative sentences from the **T-1** dataset (to be described in section 3.2) were annotated using AMT for words that most contribute to the overall sentiment of the sentence (marking words that go against the dominant sentiment was omitted to simplify the protocol). Each sentence was assigned to 5 AMT annotators; all words marked by at least 3 AMT annotators were selected.

Finally, the **Seed Lexicon** was created by adding to Lexicon 0 all words[3] marked in L-1 annotations and all words selected from the AMT annotations; the authors then performed a manual clean-up. The resulting Seed Lexicon contains 749 words, 406 positive and 343 negative. L-2 was not used in lexicon creation. However, the labeled sentences in that set were used in the evaluation experiments described in section 3.2.

2.2 Automatically Expanding the Seed Lexicon

We used three sources to automatically expand the Seed Lexicon: WordNet [3], Lin's distributional thesaurus [4], and a pivot-based paraphrase generation tool [5]. The resulting lexicons will be called Raw WN, Raw Lin, and Raw Para, respectively; they were created as follows. Please see column 2 of Table 3 for sizes of these lexicons.

Raw WN We used WordNet to extract the first three synonyms of the first sense of each word in the Seed Lexicon, restricting returned words to those with the same part-of-speech as the original word. The selection process was based on previous research [6] that showed that these constraints are likely to produce strong synonym candidates without a large number of false positives.

RAW Lin Lin's proximity-based thesaurus trained on our in-house essay data as well as on well-formed newswire texts provided an additional resource for lexicon expansion. All words with the proximity score > 1.80 to any of the Seed Lexicon words were included in the expansion.

RAW Para We used a pivot-based lexical and phrasal paraphrase generation system. This system operates by extracting bilingual correspondences from a bilingual corpus, identifying phrases and words in Language A that all correspond to the same phrase or word in Language B and pivoting on the common phrase or word to extract all Language A words and phrases as paraphrases of each other. More details can be found in [7]. We use the French-English parallel corpus (approx. 1.2 mln sentences) from the corpus of European parliamentary proceedings [8]. The base paraphrase system is susceptible to noise due to the imperfect bilingual word alignments. Therefore, we implement additional heuristics in order to minimize the number of noisy paraphrase pairs [5]. For example, one such heuristic filters out any pairs where a function word may have been inferred as a paraphrase of a content word. For lexicon expansion experiments, we use the top 15 single-word paraphrases for every word from the Seed Lexicon, excluding morphological variants of the seed word.

[3] When annotators could not attribute sentiment to single words, they marked phrases. Our current lexicons make no use of multi-word annotations.

Table 1 shows some examples for each expansion method. We note that *only* the distributional thesaurus is based on in-domain data, while the other expansions use either a general-purpose dictionary (WordNet) or out-of-domain training materials (parliamentary proceedings). It therefore remains to be seen whether the generated expansions are sufficiently general to apply to our domain.

Table 1. Examples of words added through various expansion methods.

Seed Word	WN expansion	LIN expansion	Para expansion
abuse	ill-treatment	harassment	exploitation
accuse	incriminate	indict	reproach
anxiety	anxiousness	anguish	disquiet
conflict	battle	clash	crisis
costly	dearly-won	burdensome	onerous
dangerous	unsafe	deadly	toxic
improve	amend	enhance	reinforce
invaluable	priceless	valuable	precious

2.3 Manual Clean-up of Expanded Lexicons

The expanded lexicons did require some clean-up. For example, antonyms of positive words ended up in the positive lexicon. This could happen, especially when using the distributional thesaurus, since antonyms and their synonyms tend to appear in the same distributions (for example, a *good* paper; a *bad* paper). In order to narrow down the expansions to words that carry positive or negative sentiment, we employ Amazon Mechanical Turk again. All words generated by at least one expansion mechanism were included in this new task. This time, AMT annotators were asked to label single words as positive, negative or neutral. Each word received three annotations. Two filtering criteria were used to generate the Cleaned versions based on the original raw expanded lexicon: (a) Each word had be tagged as POS or NEG polarity by the majority of the three annotators, and (b) The majority polarity had to match the expected polarity of the word, that is, the polarity of the word from the Seed Lexicon for which the current word had been generated as an expansion. The clean-up procedures resulted in 16% to 41% reduction in the Raw lexicon sizes (see Table 3 for the sizes of the Raw and Cleaned lexicons), although the Cleaned lexicons are still at least twice the size of the Seed Lexicon.

2.4 Automatic Clean-up of Expanded Lexicons through Lexicon Combination

We also experimented with an alternative strategy for noise reduction in the expanded lexicons. This strategy is based on the assumption that the three sources and mechanisms of expansion are sufficiently different to provide independent evidence of a word’s relatedness to the seed word it was expanded from. Therefore, in the generation of new lexicons, where automated clean-up was applied, we introduced into the new lexicon *only* words that were produced by both of the two expansion methods of choice. Thus, we created the Raw WN + Raw Lin lexicon that contains the Seed Lexicon expanded with words with the same polarity that were both in Raw WN and in Raw Lin. Raw WN + Raw Para lexicon and Raw Lin + Raw Para lexicon were generated in a similar fashion. This procedure resulted in the elimination of up to 63% of the larger of the two raw lexicons; for the lexicon sizes, see Table 3. Still, these new lexicons are 29%-55% larger than the Seed Lexicon, providing significant expansion without human intervention.

3 Evaluating the Quality of the Lexicons

3.1 Evaluation Methodology

We used C5.0 [9], a decision-tree classifier, to evaluate the effectiveness of the different lexicons for classifying sentiment polarity of sentences. Each lexicon is represented by just two features: (1) the number of positive words in the sentence, and (2) the number of negative words in the sentence. A number of experiments were run with additional features, but using only these two features produced the highest accuracies. For instance, an additional feature that was tried was the difference between the number of positive and negative words in a sentence. This is relevant since a sentence with a positive polarity, for instance, can contain negative words. We hypothesized that a large difference in counts might help to predict the dominant polarity. However, adding this difference feature did not boost performance.

3.2 Data Sets for Training and Testing

To generate the data for training and testing C5.0, the following strategies were employed. We used our pool of 100,000 essays to sample a second, non-overlapping set of 5,000 essays. From these essays, we randomly sampled 550 sentences, and submitted them to sentiment polarity annotation by two research assistants, both of whom had experience doing linguistic annotation. Fifty of the sentences were annotated by both annotators, with a resulting Kappa of 0.8. Sentences labeled as incomprehensible or as containing both negative and positive sentiment in equal measure were removed. The remaining sentences were randomly split between **T-1** and **TEST** sets, except for the 43 sentences out of the 50 double-annotated ones for which the annotators were in agreement. These sentences were all added to the TEST set. T-1 contains 247 sentences. TEST contains 281 sentences; this is the set used for the blind testing reported in Table 4.

As a second step, in order to augment the training set, we utilized the data initially used for lexicon development. Recall that L-1 and L-2 had been created with a bias towards positive and negative polarity sentences (see section 2.1); this resulted in a significantly smaller proportion of NEU sentences in L-1 and L-2 (11%) than in the T-1 set (39%). To mitigate the risk of a significantly different category distribution between training and testing materials, we wanted to create a larger training set that matched the distribution of T-1. We implemented the following procedure to create **T-2**. We sampled data from L-1 and L-2 such as to match the category distributions in T-1. This resulted in the utilization of all NEU sentences in L-1 and L-2, and of only a small proportion of POS and NEG sentences in these sets. Adding the new items to T-1, we now have T-2 (482 sentences), doubling the amount of training data and retaining the T-1 distribution of categories.

In order to further expand the training data without changing its category distribution, we used the remaining POS and NEG annotated sentences in L-1 and L-2 and undertook the following procedure to collect more neutral sentences. Using a different essay pool with the same sub-genres of essays, we randomly sampled 1000 sentences with a condition that was complementary to the one used to produce L-1 and L-2, that is, with the condition that *none* of the sampled sentences match any word in Lexicon 0 (see section 2.1). This way, we obtained a higher percentage of NEU sentences in the sample. This new 1000 sentence set was submitted to AMT, and all sentences with a majority vote of NEU out of 3 AMT annotator labels were considered to be acceptable neutral sentences. We then added these NEU sentences, along with the appropriate number of POS and NEG sentences from L-1 and L-2[4] to maintain the category distribution of T-1 and T-2, and produced the final training set, **T-3**. Table 2 summarizes the sizes of all three training sets and of the test set.

Table 2. Sizes of the training and test sets for c5.0 experiments. The distribution of categories is the same in all training sets, 39% NEU, 35% POS, 26% NEG.

Dataset	# Sentences
T-1	247
T-2	482
T-3	1631
TEST	281

3.3 Effective Coverage

Before moving to our evaluation that examines the performance of the family of lexicons on a sentence-level sentiment polarity classification task, we checked whether or not the expansion strategies actually succeeded in expanding the *effective coverage* of the data. Specifically, we examined whether the words added during expansion appear in our three training sets. This question is especially pertinent to our expansion strategies that used corpus-independent material (such as WordNet) or out-of-domain material, like the paraphrase generation tool. Table 3 shows the sizes of the

[4] We added 160 POS sentences from the newly annotated 1000 to T-3, in order to retain the distribution of categories the same as in T-1.

lexicons as well as the number of different words from the lexicon that were in fact observed in the T-1, T-2, and T-3 datasets. Note that there is no guarantee that an observed word is a sentiment-bearing word; performance of each lexicon in the evaluation through sentiment classification as reported in the next section addresses this aspect of the expansion process.

Table 3 shows that even the most conservative expansion (Raw WN + Raw Lin) is 29% bigger than the Seed Lexicon and has 17% more coverage than Seed Lexicon for the largest training set. We also note that both the manually- and the automatically-cleaned lexicons are on par with the state-of-art lexicons MPQA and GI in terms of effective coverage, even though they are at least 50% smaller in overall size.

Table 3. Sizes and effective sizes of the various subjectivity lexicons.

Lexicon	# Words	#Words in T-1	#Words in T-2	#Words in T-3
Seed Lexicon	749	198	390	675
Raw WN	2,527	479	867	1414
Raw Lin	1,907	292	563	981
Raw Para	2,994	585	1028	1679
Cleaned WN	1,495	280	541	936
Cleaned Lin	1,594	249	495	880
Cleaned Para	1,896	393	718	1220
Raw WN + Raw Lin	967	232	457	788
Raw WN + Raw Para	1,158	326	601	973
Raw Lin + Raw Para	1,118	246	486	836
MPQA	6,450	244	504	1014
GI	3,628	243	491	923

3.4 Prediction of Sentiment Polarity

Table 4 below summarizes the accuracy of C5.0 classifier using counts of POS and NEG words as features, across various lexicons and the three cumulatively larger training sets with the same category distributions. All systems are evaluated on the TEST set. The majority baseline – 0.466 –corresponds to the proportion of NEU cases in TEST set. For the best system (Raw WN + Raw Para, T-3, accuracy of 0.548) the following are the Precision, Recall, and F-measures for the POS, NEG, and NEU categories: POS: P = 0.56, R = 0.44, F = 0.49; NEG: P = 0.45, R = 0.34, F = 0.39; NEU: P = 0.57, R = 0.72, F = 0.64.

Table 4. Accuracy of C5.0 classification of sentiment polarity. The best 4 runs for our lexicons are bold-faced and marked with an asterisk (*).

Lexicon	T-1	T-2	T-3
Majority Baseline	0.466	0.466	0.466
Seed Lexicon	0.520	0.512	0.512
Raw WN	0.448	0.473	0.456
Raw Lin	0.452	0.452	0.466
Raw Para	0.445	0.459	0.459
Cleaned WN	0.498	0.523	0.523
Cleaned Lin	**0.537***	0.505	0.491
Cleaned Para	0.473	0.484	0.505
Raw WN + Raw Lin	**0.544***	0.505	0.491
Raw WN + Raw Para	0.498	0.530	**0.548***
Raw Lin + Raw Para	0.516	**0.537***	0.484
MPQA	0.523	0.541	0.544
GI	0.512	0.530	0.491

4 Related Work

The research into subjectivity analysis can be clustered into three primary strands: (a) identification of words that are linked to subjectivity, (b) identification of subjectivity in sentences, clauses, phrases and words in a context – that is, subjective/objective, and positive/negative polarity classification, and (c) identification of subjectivity for applications, such as determining the sentiment orientation of reviews [10-12], of news headlines [13] and articles [1,14], or of blogs [15].

We are interested in exploiting subjectivity analysis for automated essay evaluation and scoring. Specifically, the target application would have the sentiment analysis system working in tandem with a discourse analysis system [16], and allow, for example, identification of opinion orientation in the thesis statement of an issue essay, or determine the existence of both orientations in the summary statement of an essay written for the task of summarizing and evaluating contrasting opinions on a given subject. Our focus is, therefore, on sentence-level sentiment analysis, rather than phrase- or document-level, as in much of the current literature.

Identification of subjectivity in context has largely been left untouched in our work so far, and constitutes a major direction for future research. Approaches include the use of negation to alter the prior polarity of sentiment words, both grammatical negation (*not happy*) and content-word negations (*prevent further deterioration*), as well as of intensifiers (*extremely efficient* vs *somewhat efficient*), and identification of construction not representing a statement of belief, such as conditionals [17-20].

The step that concerned us most so far is the first step of compiling a comprehensive list of words with clear prior polarity ('prior' meaning before any contextual alterations occur, such as negation). Probably the broadest recent research initiative in this area is the work of Wiebe and colleagues. A concrete outcome of various annotation studies is the MPQA subjectivity lexicon, freely available for research: http://www.cs.pitt.edu/mpqa/ [1,14,21]. This lexicon contains a list of words labeled with information about polarity (positive, neutral and negative), and the intensity of the polarity (strong or weak). More recent and detailed lexicons that include word sense information are also freely available (at the website above) and are extensions of this work [22,23]. Additional lexicons include the classic General Inquirer [2], the sense-based SentiWordNet [24], as well as custom-built lexicons with wide coverage such as in [19]. In the current work, our lexicons are compared to both MPQA and GI to provide a comparative performance and coverage evaluations.

The closest related work to our current project are studies dealing with expansions of subjectivity lexicons. The most popular source for expansion, also utilized in our work, is WordNet [24-31]. Additional resources include [32], symmetric syntactic patterns like conjunctions [33], as well as distributional information derived from a large corpus (12, 34, 35). In the latter setting, words are classified based on their distributional similarity to a small seed set of positive (negative) words; our approach in using Lin's distributional thesaurus for lexicon expansion is in a similar spirit.

Over the last decade, data-driven paraphrase generation has become an extremely active area in NLP. In particular, it has been used to improve several tasks such as query expansion in Information Retrieval [36-37], evaluation of NLP systems [38-40] and statistical machine translation [41,42]. A more comprehensive review of paraphrase generation and its applications can be found in [43]. To our knowledge, ours is the first reported attempt to use a paraphrase generation system for expanding a subjectivity lexicon.

In recent years, online crowdsourcing services utilizing a scalable, anonymous workforce have emerged as cost-effective means for collecting linguistic annotations. In particular, Amazon Mechanical Turk has become a popular resource for non-expert annotation of linguistic data for use in diverse NLP applications [44-47], including sentiment analysis [48-50]. While our test data was annotated by research assistants, we elected to employ AMT at various stages of lexicon development and for generating additional training data.

5 Discussion and Conclusion

Based on Tables 2-4, the following points deserve mention.

(a) The top four runs of the expanded lexicons (see Table 4) all outperform the best run for Seed Lexicon (0.520 vs 0.548, 0.544, 0.537). These results support cautious optimism regarding the chosen lexicon expansion strategies.

(b) In the top 4 runs, one represents a lexicon that underwent manual AMT-based cleaning (Clean Lin), while the other three, including the top performer, were produced by automatically combining different expansions

(Raw WN and Raw Lin, Raw WN and Raw Para, Raw Lin and Raw Para). There is therefore some evidence that effective use of the complementarity of automated expansion methods can produce results of comparable quality to human clean-up.

(c) The best accuracy is obtained on a run with Raw WN and Raw Para, suggesting that a paraphrase generation system developed on out-of-domain data is effective for expansion of a subjectivity lexicon. To our knowledge, this is the first attempt to use a paraphrase generation system for this task, and it is shown to hold promise.

(d) The top 4 performers are on par with the best run for the state-of-art MPQA lexicon (0.544) and perform better than the General Inquirer lexicon (0.530). We therefore showed it to be feasible to build a competitive subjectivity lexicon from scratch using steps described in this paper.

(e) The better performance of combinations of raw lexicons (Raw Lin + Raw WN, etc) relative to individual raw lexicons suggests that the different expansion strategies are complementary to a certain degree. In future work, we will investigate possibilities for combining multiple lexicons.

(f) The improvement with the size of the training data is not consistent. We suspect that the reason lies in the variability in our data. The data is sampled from a large pool of materials as diverse as argumentative essays on general topics and technical essays from accounting exams. Apparently, the fit in the category distribution is not sufficient for effective utilization of the training data, as additional training items might differ substantially from the T-1 and TEST data. We might need to fit a different model to every essay type, or use a much larger sample of sentences that represents all the different sub-types. The particulars of the sampling procedure might also matter; for example, the NEU sentences added in T-2 might be more difficult than those in T-1 and TEST, since they contain distracters – sentiment words from Lexicon 0. Consider “Could there be a reason for the participants of the study to say they are eating healthier?” that uses a generally positive term *healthier* versus “To end the inning you have to get three outs.”

References

1. Wiebe, J., Wilson, T., and Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2-3), pp. 165-210 (2005).
2. Stone, P., Dunphy, D., Smith, M., Ogilvie, D.: The General Inquirer: A Computer Approach to Content Analysis. MIT (1966).
3. Miller, G.: WordNet: A lexical database. Communications of the ACM, 38(11) 39-41 (1995).
4. Lin, D.: Automatic retrieval and clustering of similar word. In: Proceedings of the ACL, Montreal, Canada (1998).

5. Madnani, N., Dorr, B.: Generating Targeted Paraphrases for Improved Translation. ACM Transactions on Intelligent Systems and Technology, 3(2) (To appear).
6. Burstein, J., Pedersen, T.: Towards Improving Synonym Options in a Text Modification Application. *University of Minnesota Supercomputing Institute Research Report Series,* UMSI 2010/165 (2010).
7. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: Proceedings of the ACL, pp. 597-604, Ann Arbor, MI (2005).
8. Koehn, P.: EUROPARL: A Parallel corpus for Statistical Machine Translation. In: Proceedings of the Machine Translation Summit (2005).
9. Quinlan, J. R.: C4.5: Programs for machine learning. Morgan Kaufmann Publishers (1993).
10. Popescu, A., Etzioni, O.: Extracting product features and opinions from reviews. In: Proceedings of HLT/EMNLP, pages 339–346 (2005).
11. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL, pp. 271–278 (2004).
12. Turney, P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the ACL, pages 417–424, Philadelphia (2002).
13. Strapparava, C. and Mihalcea, R.: SemEval-2007 Task 14: Affective Text. In: Proceedings of SemEval, Prague, Czech Republic (2007).
14. Wilson, T., Wiebe, J., and Hoffmann, P.: Recognizing Contextual Polarity in Phrase Level Sentiment Analysis. In: Proceedings of HLT-EMNLP (2005).
15. Godbole, N., Srinivasaiah, M., and Skiena, S.: Large Scale Sentiment Analysis for News and Blogs. In: Proceedings of ICWSM (2007).
16. Burstein, J., Marcu, D., and Knight, K.: Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. In S. Harabagiu and F. Ciravegna (Eds.) *Special Issue on Advances in NLP, IEEE Intelligent Systems,* 18(1) 32-39 (2003).
17. Choi, Y. and Cardie, C.: Learning with Compositional Semantics as Structure Inference for Subsentential Sentiment Analysis. In: Proceedings of EMNLP, pp. 793-801 (2008).
18. Moilanen, K. and Pulman, S.: Sentiment Composition. In: Proceedings of Recent Advances in NLP (RANLP), pp. 378–382. September, Borovets, Bulgaria (2007).
19. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M.: Lexicon-Based Method for Sentiment Analysis. Computational Linguistics 37 (2): 267-307 (2011).
20. Polanyi, L., Zaenen, A.: Contextual valence shifters. In Wiebe, ed., Computing Attitude and Affect in Text: Theory and Applications. Springer, Dordrecht, pages 1–10 (2006).
21. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of EMNLP, pages 105–112 (2003).
22. Wiebe, J. Mihalcea, R.: Word Sense and Subjectivity. In: Proceedings of ACL (2006).
23. Gyamfi, Y., Wiebe, J., Mihalcea, R., Akkaya, C.: Integrating Knowledge for Subjectivity Sense Labeling. In: Proceedings of NAACL (2009).
24. Esuli, A., Sabastiani, F.: Determining term subjectivity and term orientation for opinion mining. In: Proceedings of the EACL (2006).
25. Kim, S., Hovy, E.: Determining the sentiment of opinions. In: Proceedings of COLING (2004).
26. Andreevskaia, A., Bergler, S.: Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In: Proceedings of the EACL (2006).
27. Hu, M. Liu, B.: Mining and summarizing customer reviews. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp.168-177 (2004).
28. Kanayama, H., Nasukawa, T.: Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. In: Proceedings of EMNLP (2006).
29. Strapparava, C., Valitutti, A.: WordNet-affect: and affective extension of Word-Net. In: Proceedings of LREC, Lisbon, Portugal (2004).
30. Kamps, J., Marx, M., Mokken, R., de Rijke, M.: Using WordNet to measure semantic orientation of adjectives. In: Proceedings of LREC (2004).

31. Takamura, H., Inui, T., Okumura, M.: Extracting semantic orientation of words using spin model. In: Proceedings of the ACL, pages 133–140 (2005).
32. Mohammad, S., Dunne, C., and Dorr, B.: Generating High-Coverage Semantic Orientation Lexicons from Overtly Marked Words and a Thesaurus. In: Proceedings of EMNLP (2009).
33. Hatzivassiloglou, V., McKeown, K.: Predicting the semantic orientation of adjectives. In: Proceedings of the ACL, pages 174–181 (1997).
34. Turney, P., Littman, M.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21(4):315–346 (2003).
35. Yu, H., Hatzivassiloglou, V.: Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of EMNLP, pages 129–136, Morristown, NJ (2003).
36. Metzler, D., Dumais, S., Meek, C.: Similarity measures for short segments of text. In: Proceedings of the European Conference on Information Retrieval, pp. 16–27 (2007).
37. Riezler, S., Vasserman, A., Tsochantaridis, I., Mittal, V., Liu, Y.: Statistical machine translation for query expansion in answer retrieval. In: Proceedings of ACL (2007).
38. Zhou, L., Lin, C., Muntenau, D., Hovy, E.: ParaEval: Using paraphrases to evaluate summaries automatically. In Proceedings of HLT-NAACL, pp. 447–454 (2006).
39. Owczarzak, K., Groves, D., van Genabith, J., Way, A.: Contextual bitext-derived paraphrases in automatic MT evaluation. In Proceedings on the Workshop on Statistical Machine Translation, pages 86–93, New York (2006).
40. Kauchak, D., Barzilay, R.: Paraphrasing for automatic evaluation. In Proceedings of HLT-NAACL, pages 455–462, New York (2006).
41. Callison-Burch, C., Koehn, P., Osborne, M.: Improved statistical machine translation using paraphrases. In Proceedings of NAACL, pages 17–24, New York (2006).
42. Madnani, N., Resnik, P., Dorr, B., Schwartz, R.: Are multiple reference translations necessary? Investigating the value of paraphrased reference translations in parameter optimization. In Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA), pages 143–152, Waikiki, HI (2008).
43. Madnani, N., Dorr, B.: Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods. Computational Linguistics, 36(3):341-387 (2010).
44. Snow, R., O'Connor, B., Jurafsky, D., and Ng, Y.: Cheap and fast -- but is it good?: Evaluating non-expert annotations for natural language tasks. In: Proceedings of EMNLP. Association for Computational Linguistics, Stroudsburg, PA, USA, 254-263 (2008).
45. Sheng, V., Provost, F., and Ipeirotis, P.: Get another label? Improving data quality and data mining using multiple, noisy labelers. In: Proceeding of KDD, pp. 614-622 (2008).
46. Callison-Burch, C.: Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In: Proceedings of EMNLP, pp. 286-295 (2010).
47. Callison-Burch, C. and Dredze, M.: Creating speech and language data with Amazon's Mechanical Turk. In: Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 1-12 (2010).
48. Akkaya, C., Conrad, A., Wiebe, J., Mihalcea, R.: Amazon Mechanical Turk for subjectivity word sense disambiguation. In: Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 195-203 (2010).
49. Mellebeek, B., Benavent, F., Grivolla, J., Codina, J., Costa-juss, M., and Banchs, R.: Opinion mining of Spanish customer comments with non-expert annotations on Mechanical Turk. In: Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 114-121 (2010).
50. Mohammad, S. Turney, P.: Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In: Proceedings of NAACL Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, 26-34 (2010).