Lexical Acquisition for Opinion Inference: A Sense-Level Lexicon of Benefactive and Malefactive Events

Yoonjung Choi[1], **Lingjia Deng**[2], **and Janyce Wiebe**[1,2]
[1]Department of Computer Science
[2]Intelligent Systems Program
University of Pittsburgh
yjchoi@cs.pitt.edu, lid29@pitt.edu, wiebe@cs.pitt.edu

Abstract

Opinion inference arises when opinions are expressed toward states and events which positive or negatively affect entities, i.e., benefactive and malefactive events. This paper addresses creating a lexicon of such events, which would be helpful to infer opinions. Verbs may be ambiguous, in that some meanings may be benefactive and others may be malefactive or neither. Thus, we use WordNet to create a sense-level lexicon. We begin with seed senses culled from FrameNet and expand the lexicon using WordNet relationships. The evaluations show that the accuracy of the approach is well above baseline accuracy.

1 Introduction

Opinions are commonly expressed in many kinds of written and spoken text such as blogs, reviews, new articles, and conversation. Recently, there have been a surge in reserach in *opinion analysis* (*sentiment analysis*) research (Liu, 2012; Pang and Lee, 2008).

While most past researches have mainly addressed explicit opinion expressions, there are a few researches for implicit opinions expressed via *implicatures*. Deng and Wiebe (2014) showed how sentiments toward one entity may be propagated to other entities via opinion implicature rules. Consider *The bill would curb skyrocketing health care costs*. Note that *curb* costs is bad for the object *costs* since the costs are reduced. We can reason that the writer is positive toward the event *curb* since the event is bad for the object *health care costs* which the writer expresses an explicit negative sentiment (*skyrocketing*). We can reason from there that the writer is positive toward *the bill*, since it is the agent of the positive event.

These implicature rules involve events that positively or negatively affect the *object*. Such events are called *malefactive* and *benefactive*, or, for ease of writing, *goodFor* (*gf*) and *badFor* (*bf*) (hereafter *gfbf*). The list of gfbf events and their polarities (gf or bf) are necessary to develop a fully automatic opinion inference system. On first thought, one might think that we only need lists of gfbf *words*. However, it turns out that gfbf terms may be ambiguous – a single word may have both gf and bf meanings.

Thus, in this work, we take a sense-level approach to acquire gfbf lexicon knowledge, leading us to employ lexical resources with fine-grained sense rather than word representations. For that, we adopt an automatic bootstrapping method which disambiguates gfbf polarity at the sense-level utilizing WordNet, a widely-used lexical resource. Starting from the seed set manually generated from FrameNet, a rich lexicon in which words are organized by semantic frames, we explore how gfbf terms are organized in WordNet via semantic relations and expand the seed set based on those semantic relations.

The expanded lexicon is evaluated in two ways. First, the lexicon is evaluated against a corpus that has been annotated with gfbf information at the word level. Second, samples from the expanded lexicon are manually annotated at the sense level, which gives some idea of the prevalence of gfbf lexical ambiguity and provides a basis for sense-level evaluation. Also, we conduct the agreement study. The results show that the expanded lexicon covers more than half of the gfbf instances in the gfbf corpus, and the system's accuracy, as measured against the sense-level gold standard, is substantially higher than baseline. In addition, in the agreement study, the annotators achieve good agreement, providing evidence that the annotation task is feasible and that the concept of gfbf gives us a natural coarse-grained grouping of senses.

Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 107–112,
Baltimore, Maryland, USA. June 27, 2014. ©2014 Association for Computational Linguistics

2 The GFBF Corpus

A corpus of blogs and editorials about the *Affordable Care Act*, a controversial topic, was manually annotated with gfbf information by Deng et al. (2013)[1]. This corpus provides annotated gfbf events and the agents and objects of the events. It consists of 134 blog posts and editorials. Because the Affordable Health Care Act is a controversial topic, the data is full of opinions. In this corpus, 1,411 gfbf instances are annotated, each including a gfbf event, its agent, and its object (615 gf instances and 796 bf instances). 196 different words appear in gf instances and 286 different words appear in bf instances; 10 words appear in both.

3 Sense-Level GFBF Ambiguity

A word may have one or more meanings. For that, we use WordNet[2], which is a large lexical database of English (Miller et al., 1990). In WordNet, nouns, verbs, adjectives, and adverbs are organized by semantic relations between meanings (*senses*). We assume that a sense is exactly one of gf, bf, or neither. Since words often have more than one sense, the polarity of a **word** may or may not be consistent, as the following WordNet examples show.

- A word with only gf senses: **encourage**
  S1: (v) promote, advance, boost, further, encourage (contribute to the progress or growth of)
  S2: (v) encourage (inspire with confidence; give hope or courage to)
  S3: (v) encourage (spur on)

- A word with only bf senses: **assault**
  S1: (v) assail, assault, set on, attack (attack someone physically or emotionally)
  S2: (v) rape, ravish, violate, assault, dishonor, dishonour, outrage (force (someone) to have sex against their will)
  S3: (v) attack, round, assail, lash out, snipe, assault (attack in speech or writing)

All senses of *encourage* are good for the object, and all senses of *assault* are bad for the object. The polarity is always same regardless of sense. In such cases, for our purposes, which particular sense is being used does not need to be determined because any instance of the word will be good for (bad for); that is, word-level approaches can work well. However, word-level approaches are not applicable for all the words. Consider the following:

- A word with gf and neutral senses: **inspire**
  S3: (v) prompt, inspire, instigate (serve as the inciting cause of)
  S4: (v) cheer, root on, inspire, urge, barrack, urge on, exhort, pep up (spur on or encourage especially by cheers and shouts)
  S6: (v) inhale, inspire, breathe in (draw in (air))

- A word with bf and neutral senses: **neutralize**
  S2: (v) neutralize, neutralise, nullify, negate (make ineffective by counterbalancing the effect of)
  S6: (v) neutralize, neutralise (make chemically neutral)

The words *inspire* and *neutralize* both have 6 senses (we list a subset due to space limitations). For *inspire*, while S3 and S4 are good for the object, S6 doesn't have any polarity, i.e., it is a neutral (we don't think of inhaling air as good for the air). Also, while S2 of *neutralize* is bad for the object, S6 is neutral (neutralizing a solution just changes its pH). Thus, if word-level approaches are applied using these words, some neutral instances may be incorrectly classified as gf or bf events.

- A word with gf and bf senses: **fight**
  S2: (v) fight, oppose, fight back, fight down, defend (fight against or resist strongly)
  S4: (v) crusade, fight, press, campaign, push, agitate (exert oneself continuously, vigorously, or obtrusively to gain an end or engage in a crusade for a certain cause or person; be an advocate for)

As mentioned in Section 2, 10 words are appeared in both gf and bf instances. Since only words and not senses are annotated in the corpus, such conflicts arise. These 10 words account for 9.07% (128 instances) of all annotated instances. One example is *fight*. In the corpus instance *fight for a piece of legislation*, *fight* is good for the object, *a piece of legislation*. This is S4. However, in the corpus instance *we need to fight this repeal*, the meaning of *fight* here is S2, so *fight* is bad for the object, *this repeal*.

[1] Available at http://mpqa.cs.pitt.edu/corpora/gfbf/

[2] WordNet, http://wordnet.princeton.edu/

Thesefore, approaches for determining the gfbf polarity of an instance that are sense-level instead of word-level promise to have higher precision.

4 Lexicon Acquisition

In this section, we develop a sense-level gfbf lexicon by exploiting WordNet. The method bootstraps from a seed lexicon and iteratively follows WordNet relations. We consider only verbs.

4.1 Seed Lexicon

To preserve the corpus for evaluation, we created a seed set that is independent from the corpus. An annotator who didn't have access to the corpus manually selected gfbf words from FrameNet[3] in the light of semantic frames. The annotator found 592 gf words and 523 bf words. Decomposing each word into its senses in WordNet, there are 1,525 gf senses and 1,154 bf senses. 83 words extracted from FrameNet overlap with gfbf instances in the corpus. For independence, those words were discarded. Among the senses of the remaining words, we randomly choose 200 gf senses and 200 bf senses.

4.2 Expansion Method

In WordNet, verb senses are arranged into hierarchies, that is, verb senses towards the bottom of the trees express increasingly specific manners. Thus, we can follow *hypernym* relations to more general senses and *troponym* relations to more specific verb senses. Since the troponym relation refers to a specific elaboration of a verb sense, we hypothesized that troponyms of a synset tends to have its same polarity (i.e., gf or bf). We only consider the direct troponyms in a single iteration. Although the hypernym is a more general term, we hypothesized that direct hypernyms tend to have the the same or neutral polarity, but not the opposite polarity. Also, the *verb groups* are promising; even though the coverage is incomplete, we expect the verb groups to be the most helpful.

WordNet Similarity[4], is a facility that provides a variety of semantic similarity and relatedness measures based on information found in the WordNet lexical database. We choose Jiang&Conrath (1997) (*jcn*) method which has been found to be effective for such tasks by NLP researchers. When two concetps aren't related at all, it returns 0. The more they are related, the higher the value is retuned. We regarded words with similarity values greater than 1.0 to be similar words.

Beginning with its seed set, each lexicon (gf and bf) is expanded iteratively. On each iteration, for each sense in the current lexicon, all of its direct troponyms, direct hypernyms, and members of the same verb group are extracted and added to the lexicon for the next iteration. Similarity, for each sense, all words with above-threshold *jcn* values are added. For new senses that are extracted for both the gf and bf lexicons, we ignore such senses, since there is conflicting evidence (recall that we assume a sense has only one polarity, even if a word may have senses of different polarities).

4.3 Corpus Evaluation

In this section, we use the gfbf annotations in the corpus as a gold standard. The annotations in the corpus are at the word level. To use the annotations as a sense-level gold standard, all the senses of a word marked gf (bf) in the corpus are considered to be gf (bf). While this is not ideal, this allows us to evaluate the lexicon against the only corpus evidence available.

The 196 words that appear in gf instances in the corpus have a total of 897 senses, and the 286 words that appear in bf instances have a total of 1,154 senses. Among them, 125 senses are conflicted: a sense of a word marked gf in the corpus could be a member of the same synset as a sense of a word marked bf in the corpus. For a more reliable gold-standard set, we ignored these conflicted senses. Thus, the gold-standard set contains 772 gf senses and 1,029 bf senses.

Table 1 shows the results after five iterations of lexicon expansion. In total, the gf lexicon contains 4,157 senses and the bf lexicon contains 5,071 senses. The top half gives the results for the gf lexicon and the bottom half gives the results for the bf lexicon. In the table, *gfOverlap* means the overlap between the senses in the lexicon in that row and the gold-standard **gf** set, while *bfOverlap* is the overlap between the senses in the lexicon in that row and the gold-standard **bf** set. That is, of the 772 senses in the gf gold standard, 449 (58%) are in the gf expanded lexicon while 105 (14%) are in the bf expanded lexicon.

Accuracy (Acc) for gf is calculated as *#gfOverlap* / (*#gfOverlap* + *#bfOverlap*) and bf is calculated as *#bfOverlap* / (*#gfOverlap* + *#bfOverlap*).

[3] FrameNet, https://framenet.icsi.berkeley.edu/fndrupal/

[4] WN Similarity, http://wn-similarity.sourceforge.net/

goodFor				
	#senses	*#gfOverlap*	*#bfOverlap*	Acc
Total	4,157	449	176	0.72
WN Sim	1,073	134	75	0.64
Groups	242	69	24	0.74
Troponym	4,084	226	184	0.55
Hypernym	223	75	33	0.69
badFor				
	#senses	*#gfOverlap*	*#bfOverlap*	Acc
Total	5,071	105	562	0.84
WN Sim	1,008	34	190	0.85
Groups	255	11	86	0.89
Troponym	4,258	66	375	0.85
Hypernym	286	16	77	0.83

Table 1: Results after lexicon expansion

Overall, accuracy is higher for the bf than the gf lexicon. The results in the table are broken down by semantic relation. Note that the individual counts do not sum to the totals because senses of different words may actually be the same sense in WordNet. The results for the bf lexicon are consistently high over all semantic relations. The results for the gf lexicon are more mixed, but all relations are valuable.

The WordNet Similarity is advantageous because it detects similar senses automatically, so may provide coverage beyond the semantic relations coded in WordNet.

Overall, the verb group is the most informative relation, as we suspected.

Although the gf-lexicon accuracy for the troponym relation is not high, it has the advantage is that it yields the most number of senses. Its lower accuracy doesn't support our original hypothesis. We first thought that verbs lower down in the hierarchy would tend to have the same polarity since they express specific manners characterizing an event. However, this hypothesis is wrong. Even though most troponyms have the same polarity, there are many exceptions. For example, *protect#v#1*, which means the first sense of the verb *protect*, has 18 direct troponyms such as *cover for#v#1*, *overprotect#v#2*, and so on. *protect#v#1* is a gf event because the meaning is *"shielding from danger"* and most troponyms are also gf events. However, *overprotect#v#2*, which is one of troponyms of *protect#v#1*, is a bf event.

For the hypernym relation, the number of detected senses is not large because many were already detected in previous iterations (in general, there are fewer nodes on each level as hypernym links are traversed).

4.4 Sense Annotation Evaluation

For a more direct evaluation, two annotators, who are co-authors, independently annotated a sample of senses. We randomly selected 60 words among the following classes: 10 pure gf words (i.e., all senses of the words are classified by the expansion method, and all senses are put into the gf lexicon), 10 pure bf words, 20 mixed words (i.e., all senses of the words are classified by the expansion method, and some senses are put into the gf lexicon while others are put into the bf lexicon), and 20 incomplete words (i.e., some senses of the words are not classified by the expansion method).

The total number of senses is 151; 64 senses are classified as gf, 56 senses are classified as bf, and 31 senses are not classified. We included more mixed than pure words to make the results of the study more informative. Further, we wanted to included non-classified senses as decoys for the annotators. The annotators only saw the sense entries from WordNet. They didn't know whether the system classified a sense as gf or bf or whether it didn't classify it at all.

Table 2 evaluates the lexicons against the manual annotations, and in comparison to the majority class baseline. The top half of the table shows results when treating Anno1's annotations as the gold standard, and the bottom half shows the results when treating Anno2's as the gold standard. Among 151 senses, Anno1 annotated 56 senses (37%) as gf, 51 senses (34%) as bf, and 44 senses (29%) as neutral. Anno2 annotated 66 senses (44%) as gf, 55 senses (36%) as bf, and 30 (20%) senses as neutral. The incorrect cases are divided into two sets: *incorrect opposite* consists of senses that are classified as the opposite polarity by the expansion method (e.g., the sense is classified into gf, but annotator annotates it as bf), and *incorrect neutral* consists of senses that the expansion method classifies as gf or bf, but the annotator marked it as neutral. We report the accuracy and the percentage of cases for each incorrect case. The accuracies substantially improve over baseline for both annotators and for both classes.

In Table 3, we break down the results into gfbf classes. The *gf accuracy* measures the percentage of correct gf senses out of all senses annotated as gf according to the annotations (same as *bf accuracy*). As we can see, accuracy is higher for the bf than the gf. The conclusion is consistent with what we have discovered in Section 4.3.

By Anno1, 8 words are detected as mixed words, that is, they contain both gf and bf senses. By Anno2, 9 words are mixed words (this set includes the 8 mixed words of Anno1). Among the randomly selected 60 words, the proportion of mixed words range from 13.3% to 15%, according to the two annotators. This shows that gfbf lexical ambiguity does exist.

To measure agreement between the annotators, we calculate two measures: percent agreement and κ (Artstein and Poesio, 2008). κ measures the amount of agreement over what is expected by chance, so it is a stricter measure. Percent agreement is 0.84 and κ is 0.75.

	accuracy	% incorrect opposite	% incorrect neutral	base-line
Anno1	0.53	0.16	0.32	0.37
Anno2	0.57	0.24	0.19	0.44

Table 2: Results against sense-annotated data

	gf accuracy	bf accuracy	baseline
Anno1	0.74	0.83	0.37
Anno2	0.68	0.74	0.44

Table 3: Accuracy broken down for gfbf

5 Related Work

Lexicons are widely used in sentiment analysis and opinion extraction. There are several previous works to acquire or expand sentiment lexicons such as (Kim and Hovy, 2004), (Strapparava and Valitutti, 2004), (Esuli and Sebastiani, 2006), (Gyamfi et al., 2009), (Mohammad and Turney, 2010) and (Peng and Park, 2011). Such sentiment lexicons are helpful for detecting explicitly stated opinions, but are not sufficient for recognizing implicit opinions. Inferred opinions often have opposite polarities from the explicit sentiment expressions in the sentence; explicit sentiments must be combined with benefactive, malefactive state and event information to detect implicit sentiments. There are few previous works closest to ours. (Feng et al., 2011) build *connotation lexicons* that list words with connotative polarity and connotative predicates. Goyal et al. (2010) generate a lexicon of *patient polarity verbs* that imparts positive or negative states on their patients. Riloff et al. (2013) learn a lexicon of negative situation phrases from a corpus of tweets with hashtag "sarcasm".

Our work is complementary to theirs in that their acquisition methods are corpus-based, while we acquire knowledge from lexical resources. Further, all of their lexicons are word level while ours are sense level. Finally, the types of entries among the lexicons are related but not the same. Ours are specifically designed to support the automatic recognition of implicit sentiments in text that are expressed via implicature.

6 Conclusion and Future Work

In this paper, we developed a sense-level gfbf lexicon which was seeded by entries culled from FrameNet and then expanded by exploiting semantic relations in WordNet. Our evaluations show that such lexical resources are promising for expanding such sense-level lexicons. Even though the seed set is completely independent from the corpus, the expanded lexicon's coverage of the corpus is not small. The accuracy of the expanded lexicon is substantially higher than baseline accuracy. Also, the results of the agreement study are positive, providing evidence that the annotation task is feasible and that the concept of gfbf gives us a natural coarse-grained grouping of senses.

However, there is still room for improvement. We believe that gf/bf judgements of word senses could be effectively crowd-sourced; (Akkaya et al., 2010), for example, effectively used Amazon Mechanical Turk (AMT) for similar coarse-grained judgements. The idea would be to use automatic expansion methods to create a sense-level lexicon, and then have AMT workers judge the entries in which we have least confidence. This would be much more time- and cost-effective.

The seed sets we used are small - only 400 total senses. We believe it will be worth the effort to create larger seed sets, with the hope to mine many additional gfbf senses from WordNet.

To exploit the lexicon to recognize sentiments in a corpus, the word-sense ambiguity we discovered needs to be addressed. There is evidence that the performance of word-sense disambiguation systems using a similar coarse-grained sense inventory is much better than when the full sense inventory is used (Akkaya et al., 2009; Akkaya et al., 2011). That, coupled with the fact that our study suggests that many words are unambiguous with respect to the gfbf distinction, makes us hopeful that gfbf information may be practically exploited to improve sentiment analysis in the future.

7 Acknowledgments

This work was supported in part by DARPA-BAA-12-47 DEFT grant #12475008.

References

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity word sense disambiguation. In *Proceedings of EMNLP 2009*, pages 190–199.

Cem Akkaya, Alexander Conrad, Janyce Wiebe, and Rada Mihalcea. 2010. Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 195–203.

Cem Akkaya, Janyce Wiebe, Alexander Conrad, and Rada Mihalcea. 2011. Improving the impact of subjectivity word sense disambiguation on contextual opinion analysis. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 87–96.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December.

Lingjia Deng and Janyce Wiebe. 2014. Sentiment propagation via implicature constraints. In *Proceedings of EACL.*

Lingjia Deng, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/malefactive event and writer attitude annotation. In *Proceedings of 51st ACL*, pages 120–125.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of 5th LREC*, pages 417–422.

Song Feng, Ritwik Bose, and Yejin Choi. 2011. Learning general connotation of words using graph-based algorithms. In *Proceedings of EMNLP*, pages 1092–1103.

Amit Goyal, Ellen Riloff, and Hal DaumeIII. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of EMNLP*, pages 77–86.

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

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of COLING.*

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of 20th COLING*, pages 1367–1373.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 13(4):235–312.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.*

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Wei Peng and Dae Hoon Park. 2011. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. In *Proceedings of ICWSM.*

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of EMNLP*, pages 704–714.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-affect: An affective extension of wordnet. In *Proceedings of 4th LREC*, pages 1083–1086.