Benefactive/Malefactive Event and Writer Attitude Annotation

Lingjia Deng †, **Yoonjung Choi** ⋆, **Janyce Wiebe** †⋆
† Intelligent System Program, University of Pittsburgh
⋆ Department of Computer Science, University of Pittsburgh
†lid29@pitt.edu, ⋆{yjchoi,wiebe}@cs.pitt.edu

Abstract

This paper presents an annotation scheme for events that negatively or positively affect entities (*benefactive/malefactive events*) and for the attitude of the writer toward their agents and objects. Work on opinion and sentiment tends to focus on explicit expressions of opinions. However, many attitudes are conveyed implicitly, and benefactive/malefactive events are important for inferring implicit attitudes. We describe an annotation scheme and give the results of an inter-annotator agreement study. The annotated corpus is available online.

1 Introduction

Work in NLP on opinion mining and sentiment analysis tends to focus on explicit expressions of opinions. Consider, however, the following sentence from the MPQA corpus (Wiebe et al., 2005) discussed by (Wilson and Wiebe, 2005):

> (1) I think people are happy because Chavez has fallen.

The explicit sentiment expression, *happy*, is positive. Yet (according to the writer), the people are *negative* toward Chavez. As noted by (Wilson and Wiebe, 2005), the attitude toward Chavez is inferred from the explicit sentiment toward the event. An opinion-mining system that recognizes only explicit sentiments would not be able to perceive the negative attitude toward Chavez conveyed in (1). Such inferences must be addressed for NLP systems to be able to recognize the full range of opinions conveyed in language.

The inferences arise from interactions between sentiment expressions and events such as *fallen*, which negatively affect entities (*malefactive events*), and events such as *help*, which positively affect entities (*benefactive events*). While some corpora have been annotated for explicit opinion expressions (for example, (Kessler et al., 2010; Wiebe et al., 2005)), there isn't a previously published corpus annotated for benefactive/malefactive events. While (Anand and Reschke, 2010) conducted a related annotation study, their data are artificially constructed sentences incorporating event predicates from a fixed list, and their annotations are of the writer's attitude toward the events. The scheme presented here is the first scheme for annotating, in naturally-occurring text, benefactive/malefactive events themselves as well as the writer's attitude toward the agents and objects of those events.

2 Overview

For ease of communication, we use the terms *goodFor* and *badFor* for benefactive and malefactive events, respectively, and use the abbreviation ***gfbf*** for an event that is one or the other. There are many varieties of gfbf events, including destruction (as in *kill Bill*, which is bad for Bill), creation (as in *bake a cake*, which is good for the cake), gain or loss (as in *increasing costs*, which is good for the costs), and benefit or injury (as in *comforted the child*, which is good for the child) (Anand and Reschke, 2010).

The scheme targets clear cases of gfbf events. The event must be representable as a triple of contiguous text spans, $\langle agent, gfbf, object \rangle$. The agent must be a noun phrase, or it may be *implicit* (as in *the constituent will be destroyed*). The object must be a noun phrase.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 120–125,
Sofia, Bulgaria, August 4-9 2013.

Another component of the scheme is the **influencer**, a word whose effect is to either retain or reverse the polarity of a gfbf event. For example:

(2) Luckily Bill *didn't* **kill** him.
(3) The reform *prevented* companies from **hurting** patients.
(4) John *helped* Mary to **save** Bill.

In (2) and (3), *didn't* and *prevented*, respectively, reverse the polarity from badFor to goodFor (not killing Bill is good for Bill; preventing companies from hurting patients is good for the patients). In (4), *helped* is an influencer which retains the polarity (i.e., helping Mary to save Bill is good for Bill). Examples (3) and (4) illustrate the case where an influencer introduces an additional agent (*reform* in (3) and *John* in (4)).

The agent of an influencer must be a noun phrase or *implicit*. The object must be another influencer or a gfbf event.

Note that, semantically, an influencer can be seen as good for or bad for its object. A reverser influencer makes its object irrealis (i.e., not happen). Thus, it is bad for it. In (3), for example, *prevent* is bad for the *hurting* event. A retainer influencer maintains its object, and thus is good for it. In (4), for example, *helped* maintains the *saving* event. For this reason, influencers and gfbf events are sometimes combined in the evaluations presented below (see Section 4.2).

Finally, the annotators are asked to mark the writer's attitude towards the agents of the influencers and gfbf events and the objects of the gfbf events. For example:

(5) GOP **Attack on** Reform Is a **Fight Against** Justice.
(6) **Jettison** any reference to end-of-life counselling.

In (5), there are two badFor events: ⟨*GOP, Attack on, Reform*⟩ and ⟨*GOP Attack on Reform,Fight Against, Justice*⟩. The writer's attitude toward both agents is negative, and his or her attitude toward both objects is positive. In (6), the writer conveys a negative attitude toward *end-of-life counselling*. The coding manual instructs the annotators to consider whether an attitude of the writer is communicated or revealed in the particular sentence which contains the gfbf event.

3 Annotation Scheme

There are four types of annotations: gfbf event, influencer, agent, and object. For gfbf events, the agent, object, and polarity (goodFor or badFor) are identified. For influencers, the agent, object and effect (reverse or retain) are identified. For agents and objects, the writer's attitude is marked (positive, negative, or none). The annotator links agents and objects to their gfbf and influencer annotations via explicit IDs. When an agent is not mentioned explicitly, the annotator should indicate that it is *implicit*. For any span the annotator is not certain about, he or she can set the *uncertain* option to be true.

The annotation manual includes guidelines to help clarify which events should be annotated.

Though it often is, the gfbf span need not be a verb or verb phrase. We saw an example above, namely (5). Even though *attack on* and *fight against* are not verbs, we still mark them because they represent events that are bad for the object. Note that, Goyal et al. (2012) present a method for automatically generating a lexicon of what they call *patient polarity verbs*. Such verbs correspond to gfbf events, except that gfbf events are, conceptually, events, not verbs, and gfbf spans are not limited to verbs (as just noted).

Recall from Section 2 that annotators should only mark gfbf events that may be represented as a triple, ⟨*agent,gfbf,object*⟩. The relationship should be perceptible by looking only at the spans in the triple. If, for example, another argument of the verb is needed to perceive the relationship, the annotators should not mark that event.

(7) His uncle **left** him ***a massive amount of debt***.
(8) His uncle **left** him ***a treasure***.

There is no way to break these sentences into triples that follow our rules. ⟨*His uncle, left, him*⟩ doesn't work because we cannot perceive the polarity looking only at the triple; the polarity depends on *what* his uncle left him. ⟨*His uncle, left him, a massive amount of debt*⟩ isn't correct: the event is not bad for the debt, it is bad for *him*. Finally, ⟨*His uncle, left him a massive amount of debt, Null*⟩ isn't correct, since no object is identified.

Note that *him* in (7) and (8) are both considered benefactive semantic roles (Zúñiga and Kittilä, 2010). In general, gfbf objects are not equiva-

lent to benefactive/malefactive semantic roles. For example, in our scheme, (7) is a badFor event and (8) is a goodFor event, while *him* fills the benefactive semantic role in both. Further, according to (Zúñiga and Kittilä, 2010), *me* is the filler of the benefactive role in *She baked a cake for me*. Yet, in our scheme, *a cake* is the object of the goodFor event; *me* is not included in the annotations. The objects of gfbf events are what (Zúñiga and Kittilä, 2010) refer to as the primary targets of the events, whereas, they state, beneficiary semantic roles are typically optional arguments. The reason we annotate only the primary objects (and agents) is that the clear cases of attitude implicatures motivating this work (see Section 1) are inferences toward agents and primary objects of gfbf events.

Turning to influencers, there may be chains of them, where the ultimate polarity and agent must be determined compositionally. For example, the structure of *Jack stopped Mary from trying to kill Bill* is a reverser influencer (*stopped*) whose object is a retainer influencer (*trying*) whose object is, in turn, a badFor event (*kill*). The ultimate polarity of this event is goodFor and the "highest level" agent is Jack. In our scheme, all such chains of length N are treated as $N-1$ influencers followed by a single gfbf event. It will be up to an automatic system to calculate the ultimate polarity and agent using rules such as those presented in, e.g., (Moilanen and Pulman, 2007; Neviarouskaya et al., 2010).

To save some effort, the annotators are not asked to mark retainer influencers which do not introduce new agents. For example, for *Jack stopped trying to kill Bill*, there is no need to mark "trying." Of course, all reverser influencers must be marked.

4 Agreement Study

To validate the reliability of the annotation scheme, we conducted an agreement study. In this section we introduce how we designed the agreement study, present the evaluation method and give the agreement results. Besides, we conduct a second-step consensus study to further analyze the disagreement.

4.1 Data and Agreement Study Design

For this study, we want to use data that is rich in opinions and implicatures. Thus we used the corpus from (Conrad et al., 2012), which consists of 134 documents from blogs and editorials about a controversial topic, "the Affordable Care Act".

To measure agreement on various aspects of the annotation scheme, two annotators, who are co-authors, participated in the agreement study; one of the two wasn't involved in developing the scheme. The new annotator first read the annotation manual and discussed it with the first annotator. Then, the annotators labelled 6 documents and discussed their disagreements to reconcile their differences. For the formal agreement study, we randomly selected 15 documents, which have a total of 725 sentences. These documents do not contain any examples in the manual, and they are different from the documents discussed during training. The annotators then independently annotated the 15 selected documents.

4.2 Agreement Study Evaluation

We annotate four types of items (gfbf event, influencer, agent, and object) and their corresponding attributes. As noted above in Section 2, influencers can also be viewed as gfbf events. Also, the two may be combined together in chains. Thus, we measure agreement for gfbf and influencer spans together, treating them as one type. Then we choose the subset of gfbf and influencer annotations that both annotators identified, and measure agreement on the corresponding agents and objects.

Sometimes the annotations differ even though the annotators recognize the same gfbf event. Consider the following sentence:

(9) Obama **helped** reform **curb** costs.

Suppose the annotations given by the annotators were:

Ann 1. ⟨Obama, helped, curb⟩
⟨reform, curb, costs⟩
Ann 2. ⟨Obama, helped, reform⟩

The two annotators do agree on the ⟨*Obama, helped, reform*⟩ triple, the first one marking *helped* as a retainer and the other marking it as a goodFor event. To take such cases into consideration in our evaluation of agreement, if two spans overlap and one is marked as gfbf and the other as influencer, we use the following rules to match up their agents and objects:

- for a gfbf event, consider its agent and object as annotated;

- for an influencer, assign the agent of the influencer's object to be the influencer's object, and consider its agent as annotated and the newly-assigned object. In (9), Ann 2's annotations remain the same and Ann 1's become ⟨*Obama, helped, reform*⟩ and ⟨*reform, curb, costs*⟩.

We use the same measurement for agreement for all types of spans. Suppose A is a set of annotations of a particular type and B is the set of annotations of the same type from the other annotator. For any text span $a \in A$ and $b \in B$, the span coverage c measures the overlap between a and b. Two measures of c are adopted here.

Binary: As in (Wilson and Wiebe, 2003), if two spans a and b overlap, the pair is counted as 1, otherwise 0.

$$c_1(a, b) = 1 \quad if \quad |a \cap b| > 0$$

Numerical: (Johansson and Moschitti, 2013) propose, for the pairs that are counted as 1 by c_1, a measure of the percentage of overlapping tokens,

$$c_2(a, b) = \frac{|a \cap b|}{|b|}$$

where $|a|$ is the number of tokens in span a, and $\cap$ gives the tokens that two spans have in common. As (Breck et al., 2007) point out, c_2 avoids the problem of c_1, namely that c_1 does not penalize a span covering the whole sentence, so it potentially inflates the results.

Following (Wilson and Wiebe, 2003), treating each set A and B in turn as the gold-standard, we calculate the average F-measure, denoted $agr(A, B)$. $agr(A, B)$ is calculated twice, once with $c = c_1$ and once with $c = c_2$.

$$match(A, B) = \sum_{\substack{a \in A, b \in B, \\ |a \cap b| > 0}} c(a, b)$$

$$agr(A||B) = \frac{match(A, B)}{|B|}$$

$$agr(A, B) = \frac{agr(A||B) + agr(B||A)}{2}$$

Now that we have the sets of annotations on which the annotators agree, we use κ (Artstein and Poesio, 2008) to measure agreement for the attributes. We report two κ values: one for the polarities of the gfbf events, together with the effects of the influencers, and one for the writer's attitude toward the agents and objects. Note that, as in Example (9), sometimes one annotator marks a span as gfbf and the other marks it as an influencer; in such cases we regard *retain* and *goodfor* as the same attribute value and *reverse* and *badfor* as the same value. Table 1 gives the agr values and Table 2 gives the κ values.

		gfbf & influencer	agent	object
all annotations	c_1	0.70	0.92	1.00
	c_2	0.69	0.87	0.97
only certain	c_1	0.75	0.92	1.00
	c_2	0.72	0.87	0.98
consensus study	c_1	0.85	0.93	0.99
	c_2	0.81	0.88	0.98

Table 1: Span overlapping agreement $agr(A, B)$ in agreement study and consensus study.

	polarity & effect	attitude
all	0.97	0.89
certain	0.97	0.89

Table 2: κ for attribute agreement.

4.3 Agreement Study Results

Recall that the annotator could choose whether (s)he is certain about the annotation. Thus, we evaluate two sets: all annotations and only those annotations that both annotators are certain about. The results are shown in the top four rows in Table 1.

The results for agents and objects in Table 1 are all quite good, indicating that, given a gfbf or influencer, the annotators are able to correctly identify the agent and object.

Table 1 also shows that results are not significantly worse when measured using c_2 rather than c_1. This suggests that, in general, the annotators have good agreement concerning the boundaries of spans.

Table 2 shows that the κ values are high for both sets of attributes.

4.4 Consensus Analysis

Following (Medlock and Briscoe, 2007), we examined what percentage of disagreement is due to negligence on behalf of one or the other annotator (i.e., cases of clear gfbfs or influencers that were missed), though we conducted our consensus

study in a more independent manner than face-to-face discussion between the annotators. For annotator $Ann1$, we highlighted sentences for which only $Ann2$ marked a gfbf event, and gave $Ann1$'s annotations back to him or her with the highlights added on top. For $Ann2$ we did the same thing. The annotators reconsidered their highlighted sentences, making any changes they felt they should, without communicating with each other. There could be more than one annotation in a highlighted sentence; the annotators were not told the specific number.

After re-annotating the highlighted sentences, we calculate the agreement score for all the annotations. As shown in the last two rows in Table 1, the agreement for gfbf and influencer annotations increases quite a bit. Similar to the claim in (Medlock and Briscoe, 2007), it is reasonable to conclude that the actual agreement is approximately lower bounded by the initial values and upper bounded by the consensus values, though, compared to face-to-face consensus, we provide a tighter upper bound.

5 Corpus and Examples

Recall from in Section 4.1 that we use the corpus from (Conrad et al., 2012), which consists of 134 documents with a total of 8,069 sentences from blogs and editorials about "the Affordable Care Act". There are 1,762 gfbf and influencer annotations. On average, more than 20 percent of the sentences contain a gfbf event or an influencer. Out of all gfbf and influencer annotations, 40 percent are annotated as goodFor or retain and 60 percent are annotated as badFor or reverse. For agents and objects, 52 percent are annotated as positive and 47 percent as negative. Only 1 percent are annotated as none, showing that almost all the sentences (in this corpus of editorials and blogs) which contain gfbf annotations are subjective. The annotated corpus is available online[1].

To illustrate various aspects of the annotation scheme, in this section we give several examples from the corpus. In the examples below, words in square brackets are agents or objects, words in italics are influencers, and words in boldface are gfbf events.

1. And [it] will *enable* [Obama and the Democrats] - who run Washington - to get back to **creating** [jobs].
   (a) *Creating* is goodFor *jobs*; the agent is *Obama and the Democrats*.
   (b) The phrase *to get back to* is a retainer influencer. But, the agent span is also *Obama and the Democrats*, as the same with the goodFor, so we don't have to give an annotation for it.
   (c) The phrase *enable* is a retainer influencer. Since its agent span is different (namely, *it*), we do create an annotation for it.

2. [**Repealing** [the Affordable Care Act]] would **hurt** [families, businesses, and our economy].
   (a) *Repealing* is a badFor event since it deprives the object, *the Affordable Care Act*, of its existence. In this case the agent is *implicit*.
   (b) The agent of the badFor event *hurt* is the whole phrase *Repealing the Affordable Care Act*. Note that the agent span is in fact a noun phrase (even though it refers to an event). Thus, it doesn't break the rule that all agent gfbf spans should be noun phrases.

3. It is a moral obligation to *end* this indefensible **neglect of** [hard-working Americans].
   (a) This example illustrates a gfbf that centers on a noun (*neglect*) rather than on a verb.
   (b) It also illustrates the case when two words can be seen as gfbf events: both *end* and *neglect of* can be seen as badFor events. Following our specification, they are annotated as a chain ending in a single gfbf event: *end* is an influencer that reverses the polarity of the badFor event *neglect of*.

6 Conclusion

Attitude inferences arise from interactions between sentiment expressions and benefactive/malefactive events. Corpora have been annotated in the past for explicit sentiment expressions; this paper fills in a gap by presenting an annotation scheme for benefactive/malefactive events and the writer's attitude toward the agents and objects of those events. We conducted an agreement study, the results of which are positive.

Acknowledgement This work was supported in part by DARPA-BAA-12-47 DEFT grant #12475008 and National Science Foundation grant #IIS-0916046. We would like to thank the anonymous reviewers for their helpful feedback.

[1] http://mpqa.cs.pitt.edu/

References

Pranav Anand and Kevin Reschke. 2010. Verb classes as evaluativity functor classes. In *Interdisciplinary Workshop on Verbs. The Identification and Representation of Verb Features*.

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

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 2683–2688, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Alexander Conrad, Janyce Wiebe, Hwa, and Rebecca. 2012. Recognizing arguing subjectivity and argument tags. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, ExProM '12, pages 80–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Amit Goyal, Ellen Riloff, and Hal Daum III. 2012. A computational model for plot units. *Computational Intelligence*, pages no–no.

Richard Johansson and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3).

Jason S. Kessler, Miriam Eckert, Lyndsay Clark, and Nicolas Nicolov. 2010. The 2010 icwsm jdpa sentiment corpus for the automotive domain. In *4th Int'l AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Karo Moilanen and Stephen Pulman. 2007. Sentiment composition. In *Proceedings of RANLP 2007*, Borovets, Bulgaria.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2010. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 806–814, Stroudsburg, PA, USA. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):164–210.

Theresa Wilson and Janyce Wiebe. 2003. Annotating opinions in the world press. In *Proceedings of the 4th ACL SIGdial Workshop on Discourse and Dialogue (SIGdial-03)*, pages 13–22.

Theresa Wilson and Janyce Wiebe. 2005. Annotating attributions and private states. In *Proceedings of ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*.

F. Zúñiga and S. Kittilä. 2010. Introduction. In F. Zúñiga and S. Kittilä, editors, *Benefactives and malefactives*, Typological studies in language. J. Benjamins Publishing Company.