Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Acquiring Knowledge of Affective Events from Blogs Using Label Propagation

Haibo Ding and **Ellen Riloff**
School of Computing
University of Utah
Salt Lake City, UT 84112
{hbding, riloff}@cs.utah.edu

Abstract

Many common events in our daily life affect us in positive and negative ways. For example, going on vacation is typically an enjoyable event, while being rushed to the hospital is an undesirable event. In narrative stories and personal conversations, recognizing that some events have a strong affective polarity is essential to understand the discourse and the emotional states of the affected people. However, current NLP systems mainly depend on sentiment analysis tools, which fail to recognize many events that are implicitly affective based on human knowledge about the event itself and cultural norms. Our goal is to automatically acquire knowledge of stereotypically positive and negative events from personal blogs. Our research creates an event context graph from a large collection of blog posts and uses a sentiment classifier and semi-supervised label propagation algorithm to discover affective events. We explore several graph configurations that propagate affective polarity across edges using local context, discourse proximity, and event-event co-occurrence. We then harvest highly affective events from the graph and evaluate the agreement of the polarities with human judgements.

Introduction

We experience many events in our daily lives that affect us in positive and negative ways. Sometimes we express our feelings about an event using emotional words, such as *"Just graduated, I'm **so happy**!"* or *"Got laid off, **really bummed**"*. But many events are easily understood to be positive (desirable) or negative (undesirable) even if we do not explicitly express an emotion. For example, if Mary tells John that she just graduated from college, John will typically congratulate her. Conversely, if Mary tells John that she just got laid off, John will likely offer consolation.

We will refer to events that are typically associated with a positive or negative emotional state as *affective events*. Recognizing affective events and their polarity is essential for many natural language processing tasks. Understanding the emotional states of people is important for narrative story comprehension (e.g., (Lehnert 1981)), response generation (e.g., (Ritter, Cherry, and Dolan 2011)), sarcasm detection (e.g., affective dissonance (Riloff et al. 2013)), and other applications related to sentiment analysis.

The goal of our research is to acquire knowledge of common affective events from a large collection of personal blogs. While some events can be identified as positive or negative by virtue of their lexical semantics (e.g., a "celebration" is inherently good and a "catastrophe" is bad) or by identifying emotional contexts, our aim is to acquire sets of events that are implicitly affective due to world knowledge and cultural norms. For example, being "rushed to the hospital", "hit by a car", and "calling 911" describe emergency situations which indicate that the experiencer is in a negative state. Similarly, "watching a sunset", "playing games", and "having a cookout" are recreational activities which indicate that the experiencer is in a positive state. Existing sentiment analysis tools do not recognize many of these events as being affective because the individual words do not carry sentiment. Our research automatically identifies and harvests affective events from a large text corpus, toward the long-term goal of creating a resource of affective event knowledge.

To identify events that are typically associated with a positive or negative state, we explore semi-supervised learning with graphs using a large collection of personal blogs. First, we extract all of the events in the blog posts using a shallow event representation. We then construct an enormous *Event Context Graph* that contains nodes representing the events and sentences in the corpus. Our approach is based on the intuition that some instances of affective events will occur in emotional contexts, so the graph is designed to link event mentions with their contexts. We explore graph configurations that incorporate three types of contexts as edges: *event-sentence* edges capture local context, *sentence-sentence* edges capture discourse proximity context, and *event-event* edges capture event co-occurrence in documents. We then assign initial affective polarities with a sentiment classifier to "seed sentences" and use a semi-supervised label propagation algorithm to spread affective evidence across edges of the graph. Our results show that graph-based label propagation learns to identify many stereotypically positive and negative affective events with good accuracy.

Related Work

NLP researchers have been studying the problem of constructing lexicons and knowledge bases for sentiment analysis (Pang and Lee 2008). Many sentiment-related resources

have been created, including the MPQA Subjectivity Lexicon (Wilson, Wiebe, and Hoffmann 2005), SenticNet (Cambria, Olsher, and Rajagopal 2014; Cambria et al. 2015), SentiWordNet (Baccianella, Esuli, and Sebastiani 2010), ConnotationWordNet (Kang et al. 2014; Feng et al. 2013)), and others (e.g., (Hu and Liu 2004)). Additional work has focused on identifying phrases that express opinions about commercial products (e.g., (Zhang and Liu 2011; Li et al. 2015; Stone and Hunt 1963)), and learning words, phrases, and hashtags that represent specific emotions (e.g., (Mohammad and Turney 2010; Qadir and Riloff 2014)).

There has been recent research demonstrating the need to recognize affective events for specific NLP tasks and acquiring affective knowledge for those tasks. Goyal et al. (2013; 2010) tackled the problem of automatically generating plot unit representations (Lehnert 1981) and discovered that many affect states originated from affective events. For this research, they used two bootstrapping techniques to identify verbs that impart positive or negative polarity on their patients (*Patient Polarity Verbs*). Recent work on sarcasm recognition (Riloff et al. 2013) also recognized that sarcasm can arise from the juxtaposition of a positive sentiment with a negative situation (event or state). They developed a bootstrapping algorithm that automatically acquires negative situation phrases from sarcastic tweets.

Several recent research efforts have addressed problems related to recognizing affective events. Deng et al. (2013) created a manually annotated corpus of +/-effect events[1] along with the writer's attitude toward the event's agents and objects. The polarity of these events is based on the object of the event (e.g., "increased" is a +effect event because more of the object is created, so the object is affected in a positive way). Subsequent work with +/-effect events (Deng and Wiebe 2014; Deng, Wiebe, and Choi 2014; Deng and Wiebe 2015) has used them to identify positive/negative opinions toward entities and events in specific contexts. In contrast, our goal is to acquire general knowledge about stereotypically positive or negative events for the person experiencing the event (e.g., watching a sunset or having a cookout).

Recent work on "major life event" extraction (Li et al. 2014) collected major life events from tweets using a bootstrapping approach which was initialized with replies from two types of speech acts: congratulations and condolences. The tweets were clustered with topic models and then manually filtered and labeled by a human. Classifiers were then trained to label tweets with respect to 42 event categories, and to extract properties from them (e.g., name of spouse for wedding events). Vu et al.'s research (2014) acquired "emotion-provoking events" using bootstrapped learning applied to tweets with the pattern "I am EMOTION that EVENT" and seed words for six emotions. The evaluation looked at only the top 20 extracted events for five emotions. The SemEval-2015 Task 9 (Russo, Caselli, and Strapparava 2015) involves detecting the implicit polarity of events. However, these annotations of event polarity are based on specific instances of an event in context.

Label propagation algorithms have been used in previous work on sentiment lexicon induction, with graphs that define edges to exploit thesaurus relations (Rao and Ravichandran 2009), predicate-argument relations (Feng et al. 2013), and context similarity (Velikovich et al. 2010). Previous work usually initializes the label propagation algorithm with manually selected seed words. Our work initializes the graph using a sentiment classifier, and incorporates several types of edges based on discourse context.

Acquiring Affective Events with Event Context Graphs and Label Propagation

Bloggers often write about events in their daily lives. While many of these events are mundane, blog posts are often motivated by exciting events such as a vacation or graduation, or by unpleasant events such as an injury or job loss. Our goal is to learn to identify *affective events* that are stereotypically positive or negative experiences. Our approach explores the idea of harvesting affective events from a large collection of blog posts by identifying events that frequently occur in positive or negative discourse contexts. Most events, however, have neutral polarity because they describe ordinary events that are not associated with any emotional state. Consequently, a key challenge of this research is to explore the effectiveness of different types of discourse context in learning to recognize events that have a strong affective polarity.

Our approach begins by extracting frequent events from a large set of personal stories on blogs. We create an Event Context Graph that has event nodes and sentence nodes, with edges between each event node and the sentences in which it occurs. We apply a sentiment classifier to identify sentences that have strong positive or negative polarity, which become the seed nodes for semi-supervised learning. A label propagation algorithm then iteratively spreads affective evidence across the edges of the graph. Consequently, events that frequently occur in affective contexts will be assigned high values for affective polarity. We also incorporate edges between adjacent sentences to explore the benefits of spreading affective evidence across local discourse regions and we create edges to link events that co-occur in the same story. Intuitively, our hypothesis is that if two events frequently co-occur in a blog post, then they are likely to have the same affective polarity.

In the following sections, we present the technical details of our approach. First, we describe the sentiment classifier used to initialize seed nodes. Second, we explain how we represent and extract events from blog posts. Third, we present three configurations of Event Context Graphs and describe the semi-supervised label propagation algorithm.

Sentiment Sentence Classifier

We created a logistic regression classifier that labels sentences as having *positive*, *negative*, or *neutral* polarity. This classifier provides a probability value that we use to assign a strength to each label. The feature set for this classifier was modeled after the features used by the NRC-Canada sentiment classifier (Mohammad, Kiritchenko, and Zhu 2013),

[1]These events are also called goodFor/badFor or benefactive/malefactive events in different publications.

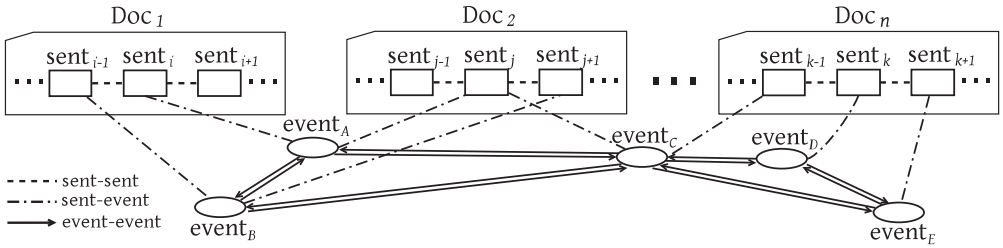

Figure 1: Illustration of an Event Context Graph with Three Types of Edges

which performed very well in the SemEval 2013 Task 2 (Sentiment Analysis in Twitter). Since our blog data is also a form of social media text, we felt that this feature set was also well-suited for our data.

The classifier's features include word n-grams, character n-grams, capitalization, part-of-speech tags, and hashtags. In addition, features are created to capture information from six sentiment lexicons: MPQA Subjectivity lexicon (Wilson, Wiebe, and Hoffmann 2005), Hu & Liu's lexicon (Hu and Liu 2004), NRC Emotion lexicon (Mohammad and Turney 2010), NRC Hashtag lexicon (Mohammad, Kiritchenko, and Zhu 2013), Sentiment140 lexicon (Go, Bhayani, and Huang 2009), and AFINN lexicon (Nielsen 2011). These features were designed to be similar to those in (Mohammad, Kiritchenko, and Zhu 2013).

We trained a logistic regression classifier with the training data from the SemEval 2014 Task 9 (Sentiment Analysis in Twitter). To evaluate its performance, we trained the classifier on 6425 of the annotated tweets and tested it on the remaining 1564 tweets. The classifier's macro average performance is 69.4% Precision, 68.2% Recall and 67.8% F1, which is similar to the results reported for the NRC-Canada system for the SemEval 2013 Message-level Task.

Event Representation and Extraction

We extract events using a shallow representation of *event triples*, which capture a verb, its agent, and the object of the verb (usually its theme). We apply the Stanford dependency parser (de Marneffe, MacCartney, and Manning 2006) to our blog collection, and extract two sets of dependency relations. For active voice verb phrase constructions, we use the *nsubj* and *dobj* relations to extract the heads of the verb's subject and its direct object. And we normalize each event by lemmatizing the subject, direct object, and the verb. Example (a) in Figure 2 would produce the event triple $\langle they, have, party \rangle$.

For passive voice verb phrase constructions, we extract the heads of the verb's subject and its agent with the *nsubjpass* and *agent* relations, and lemmatize them. Example (b) in Figure 2 shows a sentence in the passive voice, which produces the event triple $\langle man, be\_killed, by\_police \rangle$ [2].

In cases when an active voice construction does not have a direct object, or a passive voice construction does not have an agent, one of these elements may be absent (we use ϕ to indicate a missing item). But each event triple must have a verb and at least one other element. If the verb is modified by a negator, we extract the negator as well using the neg relation. Since we are interested in events, we compiled a list of 45 verbs[3] that correspond to private states and we do not create event triples from these verbs, for example: *believe*, *doubt*, *hate*, *feel*, *know*, etc.

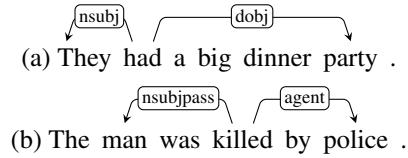

Figure 2: Dependency relations for events

Event Context Graphs

To identify events that have affective polarity, we construct an *Event Context Graph* that links events with the contexts in which they occur. We build a graph $G = (V, E)$ that contains two types of vertices (nodes): event nodes and sentence nodes. We create three kinds of graph configurations to investigate different ways of propagating affective evidence between events and contexts. We incrementally incorporate three types of edges: (a) *event-sentence* edges, (b) *sentence-sentence* edges, and (c) *event-event* edges. We describe each of the three graph architectures below.

Local Context Graph (G^{LOC}) This graph configuration contains only one type of edge: *event-sentence* edges. These edges connect event triples with the sentences in which they appear. For this graph, the affective state of an event can be induced only from its local sentential contexts. When a sentence node s_i is linked to an event node e_j, the weight is computed as $w(s_i, e_j) = \frac{1}{|T(s_i)|}$ where the $T(s_i)$ denotes the set of events linked to sentence s_i.

Discourse Context Graph (G^{DIS}) This graph configuration contains two types of edges: *event-sentence* edges as well as *sentence-sentence* edges. The *sentence-sentence* edges link adjacent sentences in the same document. They allow for label propagation across neighboring sentences to

[2] To make it easily readable, we keep the verb in its past tense and append *"be"*, and we append *"by"* to the agent.

[3] We used the list of 43 stative verbs from http://www.perfect-english-grammar.com/support-files/stative-verbs-list.pdf, and added "be" and "think".

capture the intuition that sentences in the same discourse region are likely to have the same polarity. We set the weight for an edge linking sentence s_i and s_j to be $w(s_i, s_j) = 0.80$[4], which indicates that we expect adjacent sentences to usually, but not always, have the same affective polarity.

Event Co-Occurrence Graph (G^{EV}) This graph configuration contains three types of edges: *event-sentence*, *sentence-sentence*, and *event-event* edges. The *event-event* edges link events that co-occur in the same blog post. These edges are designed to capture the intuition that if two events frequently co-occur in the same story, then they are likely to share the same affective polarity. For example, we would expect the events $\langle kid, be\_hurt, \phi \rangle$ and $\langle kid, cry, \phi \rangle$ to frequently co-occur in blog posts by parents discussing accidents involving their children.

We use the probability of event e_j given event e_i as the edge weight, so the edges are directed. The weight on an edge from event e_i to e_j is computed as:

$$w(e_i, e_j) = \frac{p(e_i, e_j)}{p(e_i)}.$$

Figure 1 shows an illustration of an Event Context Graph with *event-sentence*, *sentence-sentence*, and *event-event* edges. All of these graphs were populated with event triples that have frequency ≥ 50 in our data set, to keep the size of the graphs manageable. This produced 40,608 event nodes. In total, the Local Context Graph contained roughly 12 million nodes, and the Discourse Graph and Event Co-Occurrence graphs each contained roughly 25 million nodes.

Semi-Supervised Label Propagation

To induce the affective polarities of events, we use the label propagation algorithm from (Zhu and Ghahramani 2002). The pseudocode for our implementation is shown in Algorithm 1.

Algorithm 1 Semi-Supervised Label Propagation

Input: $G(V, E)$, Sentiment Classifier SC, seed threshold τ
Output: $\psi(v) \in [-1, +1], \forall v \in V$

1: Initialize seed nodes using SC with threshold τ
2: **while** ψ has not converged **do**
3: Update $\psi(v)$ using Equation (1).
4: Re-clamp the seed nodes
5: **end while**
6: **return** ψ

An important part of the label propagation algorithm is the initialization procedure (Step 1). Previous work (Rao and Ravichandran 2009; Feng et al. 2013) using label propagation for sentiment lexicon induction typically start with a small set of manually selected seed words. Instead, we initialize the label propagation algorithm with a set of contexts that are determined to be positive or negative. We apply a sentence classifier designed for sentiment analysis to all of the sentences in the blogs and identify sentences that are classified as having positive or negative polarity.

However, the sentiment classifier is not perfect (69% precision), and we want the initial set of labeled nodes to be as accurate as possible. So we only assign polarity to a sentence node if the classifier's probability is $\geq \tau$. The *seed nodes* correspond to the sentences that are classified as having positive or negative polarity with probability $\geq \tau$[5] by the sentiment classifier in the initialization step of Algorithm 1. The seed nodes labeled as positive are assigned a value of $+1$, and the negative seed nodes are assigned a value of -1. All other sentence nodes, and all of the event nodes, are initialized with a value of 0.

$$\psi^{t+1}(v) = \frac{\sum_{v' \in N(s)} w(v, v') * \psi^t(v')}{\sum_{v' \in N(s)} w(v, v')} \qquad \text{(Eq. (1))}$$

After initialization, the affective polarity values are iteratively propagated across edges. We compute the affective polarity $\psi(v)$ of node v as the weighted average of the polarity values of its neighbor nodes $N(v)$. Formally, we use Equation (1) to update the value of each node. After each iteration, the affective polarity value for each seed node is reset ("re-clamped") to +1 or -1, per its original value. This step ensures that the seed nodes always maintain their original polarity. The label propagation process iterates until the affective polarity values in the graph converge. For our experiments, we ran label propagation until the values converged or it ran for 100 iterations.

Evaluation

For our research, we used the *personal story corpus* compiled by Gordon & Swanson (2008), who created a system to identify stories that "are primarily a first person description of events in the life of the author". They applied their system to 44 million texts from the ICWSM 2009 Spinn3r data set, resulting in 1.4 million "personal story" texts. However many of the texts came from Web domains such as *craigslist.org* and *answers.yahoo.com* and are not narrative stories, so we extracted only the texts originating from six well-known blogging sites: livejournal.com, wordpress.com, blogspot.com, spaces.live.com, typepad.com, and travelpod.com. We removed near-duplicate entries using SpotSigs (Theobald, Siddharth, and Paepcke 2008), resulting in a text collection of 872,805 personal blog posts. We then tokenized, part-of-speech tagged, and parsed them using Stanford CoreNLP tools (Manning et al. 2014).

Baselines

For comparison, we designed two baseline systems that try to acquire affective events by applying a sentiment classifier to the contexts surrounding the events. The first system, **AvgSent**, is designed to identify affective events that frequently occur with an explicitly expressed emotion or nearby sentiment. For example, the event $\langle we, have, campfire \rangle$ is typically a fun experience, so we

[4]We chose this value based on intuition and did not experiment with other values, so exploring methods to find an optimal weighting could be fruitful.

[5]In our experiments we set $\tau = 0.5$.

expect to find sentences such as *"We will have a campfire, so excited!"*. For each event triple e, the **AvgSent** system collects all of the sentences containing the event and applies the sentiment classifier to those sentences. An event's affective polarity score $\psi(e)$ is computed as the average polarity over the sentences:

$$\psi(e) = \frac{1}{|S(e)|} \sum_{s \in S(e)} p(s)$$

where $S(e)$ is the set of sentences containing event e, $p(s)$ is the signed polarity score of sentence s which is defined as $+1$ if s is classified as positive, -1 if s is negative and 0 if s is neutral.

The second system, **AvgDoc**, is designed to identify affective events that frequently occur in documents that have an overall positive or negative polarity. For example, bloggers often express an overall sentiment at the beginning of their post (e.g. *"So happy today!"*), after which they describe the events that happened on that day. While not every event that occurs in a positive (or negative) document will have positive (or negative) polarity, if an event consistently occurs in documents with one polarity, then the event is likely to have that polarity. For each event triple e, the **AvgDoc** system collects all of the documents containing the event and applies the sentiment classifier to all sentences in those documents. Each document's sentiment score is computed as the average sentiment score for the sentences in that document. An event's affective polarity score $\psi(e)$ is then computed as the average polarity over these documents:

$$\psi(e) = \frac{1}{|D(e)|} \sum_{d \in D(e)} (\frac{1}{|d|} \sum_{s \in d} p(s))$$

where $D(e)$ is the set of documents containing event e.

Evaluation Details

We evaluated the affective events produced by the AvgSent and AvgDoc baseline systems as well as the three configurations of our Event Context Graph (G^{LOC}, G^{DIS}, G^{EV}) with label propagation. After extracting the event triples in the personal blogs data set, we applied each method and ranked the events based on the affective polarity scores produced by that method.[6]

However, many of the top-ranked events included individual words with a strong positive or negative sentiment, such as *celebrate* or *disappoint*. Events that include explicitly positive or negative terms can usually be assigned an affective polarity by sentiment analysis tools. Consequently, we applied the sentiment classifier to each event triple and separated out the events that the classifier labeled as positive or negative. We removed these cases for two reasons. First, the sentiment classifier was used to "seed" the label propagation algorithm, so it seemed unfair to reward that method for finding events that the classifier itself recognized as having polarity.[7] Second, the goal of our research was to learn *implicitly* affective events. Since manual annotation is expensive, we wanted to focus our annotation efforts on evaluating the quality of the events hypothesized to have affective polarity that would have been labeled neutral by current sentiment analysis systems.

Finally, rather than fixing an arbitrary threshold for the polarity scores, we evaluated the precision of the top-ranked k affective events hypothesized by each method. For each of the five systems, we collected the 500 top-ranked events (that were not labeled as positive or negative by the sentiment classifier). In total, this process produced 1,020 unique events, which were then assigned gold standard affective polarity labels by human annotators.

Gold Standard Annotation We used Amazon's Mechanical Turk (AMT) service to obtain gold standard annotations for the affective polarity of events. AMT workers were asked to assign one of four labels to an event triple:

Positive: the event is typically desirable or beneficial. For example: $\langle I, see, sunset \rangle$

Negative: the event is typically undesirable or detrimental. For example: $\langle girl, have, flu \rangle$

Neutral: the event is not positive or negative, or the event is so general that it could easily be positive or negative in different contexts. For example: $\langle he, open, door \rangle$

Invalid: the triple does not describe a sensible event. This label is primarily for erroneous event triples resulting from pre-processing or parsing mistakes. For example: $\langle cant, do, \rangle$

We gave annotation guidelines and examples to three AMT workers, who then annotated the 1,020 event triples. We measured pairwise inter-annotator agreement (IAA) among the three workers using Cohen's kappa (κ). Their IAA scores were κ=.73, κ=.71, and κ=.68. We assigned the majority label as the gold standard for each event triple. However, 17 event triples were assigned three different labels by the judges, and 43 event triples were annotated as Invalid events (based on the majority label), so we discarded these 60 cases. Consequently, our gold standard annotations consisted of 960 labeled event triples, which had the following distribution of affective polarities: Negative=565, Positive=198, Neutral=197. As a result of the discarded cases, we were left with less than 500 labeled events for some methods. All five methods had at least 460 labeled events, though, so we evaluated the precision of each method for its top-ranked 100, 200, 300, 400, and 460 event triples.

Systems	Top100	Top200	Top300	Top400	Top460
AvgDoc	75.0	73.5	73.3	71.5	71.5
AvgSent	86.0	83.0	83.6	82.5	80.0
G^{LOC}	88.0	86.0	84.0	81.5	80.9
G^{DIS}	88.0	87.0	**84.3**	83.0	82.4
G^{EV}	**90.0**	**87.5**	84.0	**84.5**	**82.8**

Table 1: Precision for the top-ranked affective events.

Experimental Results

Table 1 shows the accuracy of the top-ranked events produced by each system. The AvgSent baseline system performed well, achieving 86% accuracy for the top 100 documents and 80% accuracy for all 460 events. AvgDoc did

[6] We used the absolute values of the polarity scores to generate a single ranking of events with both positive and negative polarity.

[7] The sentences containing these events are likely to have been seed nodes.

Negative Events				
⟨he, lose, mind⟩	⟨she, lose, lot⟩	⟨i, break, nose⟩	⟨phone, be_broken, ∅⟩	⟨professional, advise, ∅⟩
⟨life, lose, ∅⟩	⟨she, lose, balance⟩	⟨he, lose, balance⟩	⟨phone, break, ∅⟩	⟨im, stick, ∅⟩
⟨tear, sting, eye⟩	⟨she, be_hit, by_car⟩	⟨i, fall, bike⟩	⟨he, lose, lot⟩	⟨she, hit, head⟩
⟨im, screw, ∅⟩*	⟨nose, be_stuffed,⟩	⟨he, be_hit, by_car⟩	⟨one, answer, phone⟩	⟨i, lose, balance⟩
⟨i, lose, phone⟩	⟨it, leave, taste⟩	⟨i, be_hit, by_car⟩	⟨i, twist, ankle⟩	⟨i, break, toe⟩
⟨he, lose, control⟩	⟨im, tire, ∅⟩	⟨he, have, seizure⟩	⟨neck, start, ∅⟩	⟨i, sprain, ankle⟩
⟨i, cut, finger⟩	⟨he, hit, head⟩	⟨heart, start, pound⟩	⟨he, lose, her⟩	⟨she, stick, head⟩
⟨she, lose, track⟩	⟨head, pound, ∅⟩	⟨she, lose, control⟩	⟨bone, be_broken,⟩	⟨he, lose, job⟩
⟨i, seethe, ∅⟩	⟨i, break, bone⟩	⟨she, stick, hand⟩	⟨i, screw, thing⟩*	⟨i, injure, myself⟩
⟨time, lose, ∅⟩	⟨she, black, ∅⟩*	⟨i, be_stung, by_bee⟩	⟨im, lose, ∅⟩	⟨he, be_rushed, ∅⟩
⟨nose, run, ∅⟩	⟨i, call, vet⟩	⟨she, lash, ∅⟩	⟨spell, be_broken, ∅⟩	⟨body, shut, ∅⟩*
Positive Events				
⟨we, sing, birthday⟩	⟨all, have, weekend⟩	⟨everyone, have, weekend⟩	⟨learn, make, money⟩	⟨learn, use, ∅⟩
⟨time, be_had, by_all⟩	⟨i, learn, deal⟩	⟨you, have, weekend⟩	⟨we, make, team⟩	⟨car, be_offered,⟩
⟨we, find, deal⟩	⟨it, have, view⟩	⟨we, have, turnout⟩	⟨you, have, birthday⟩	⟨kid, have, time⟩
⟨we, spend, deal⟩	⟨it, make, story⟩	⟨weather, stay, ∅⟩	⟨time, be_had, ∅⟩	⟨everyone, have, time⟩
⟨we, have, playing⟩	⟨we, have, evening⟩	⟨room, have, view⟩	⟨we, get, view⟩	⟨we, get, laugh⟩
⟨we, relax, bit⟩	⟨it, take, deal⟩	⟨we, have, view⟩	⟨we, have, weekend⟩	⟨we, get, deal⟩
⟨we, have, visit⟩	⟨practice, make, ∅⟩	⟨we, have, shopping⟩	⟨we, have, dancing⟩	⟨we, see, view⟩
⟨i, see, sunset⟩	⟨we, have, time⟩	⟨kid, have, lot⟩	⟨we, reunite, ∅⟩	⟨you, go, girl⟩
⟨i, find, deal⟩	⟨we, laugh, lot⟩	⟨we, have, turn⟩	⟨all, have, time⟩	⟨god, have, sense⟩
⟨we, have, weather⟩	⟨it, entertain, ∅⟩	⟨we, have, meal⟩	⟨we, have, cookout⟩	⟨that, give, view⟩
⟨me, motivate, ∅⟩	⟨we, have, afternoon⟩	⟨we, have, feast⟩	⟨i, get, present⟩	⟨girl, have, time⟩

Table 2: Top 55 positive and 55 negative affective events produced with label propagation with G^{EV}, ∅ denotes empty element. Verbs that usually occur with a particle are denoted with ∗ (e.g. *screw up*, *black out*, *shut down*)

not perform as well, suggesting that local sentential context is a more reliable indicator of affective polarity than document-wide context. Label propagation with the Event Context Graphs yielded additional performance gains over the AvgSent baseline. The G^{LOC} graph with only *event-sentence* edges improved precision from 83% to 86% for the top 200 events, and from 80.0% to 80.9% over all 460 events. The G^{DIS} graph with added *sentence-sentence* edges further improved precision over G^{LOC} from 80.9% to 82.4% for all 460 events. Finally, the G^{EV} graph that incorporated additional *event-event* edges achieved 90% precision for the top 100 events and slightly higher precision (82.8%) overall.

These results show that label propagation with Event Context Graphs is an effective method for acquiring affective events from a large text corpus. This approach achieved high precision at identifying affective events, and successfully discovered 380 affective events that a sentiment classifier did not recognize as having polarity. In the next section, we analyze these results further and present examples of the learned affective event knowledge.

Analysis

Table 2 shows the top 55 positive events and the top 55 negative events produced by label propagation with the G^{EV} graph. Negative events include many physical injuries and ailments, car accidents, and lost or broken phones. Note that ⟨i, call, vet⟩, ⟨she, black, ∅⟩, and ⟨he, be_rushed, ∅⟩ often suggest medical emergencies (i.e., "she blacked out" and "he was rushed to the hospital"). In future work, we plan to make the event representation richer to more precisely characterize these types of events. Positive events include birthdays, playing, dancing, shopping, and cookouts. We also see more subtle examples of stereotypically enjoyable situations, such as ⟨we, find, deal⟩, ⟨we, make, team⟩, ⟨i, see, sunset⟩, ⟨we, reunite, ∅⟩, and ⟨room, have, view⟩.

However, not all of these events are truly affective. A common source of errors are expressions that typically occur with positive/negative adjectives modifying the direct object. For example ⟨we, have, weather⟩ is not a positive or negative event per se, but originates from sentiments expressed about the weather, such as "we have nice weather". Similarly, ⟨it, leave, taste⟩ isn't negative per se, but comes from the common expression "it leaves a bad taste".

As a reminder, none of these events were identified as having positive or negative polarity by our sentiment classifier. However we further explored whether sentiment lexicons can recognize the affective polarity of these events. We used four well-known sentiment/opinion lexicons: Connotation-WordNet (Kang et al. 2014), MPQA Subjectivity Lexicon (Wilson, Wiebe, and Hoffmann 2005), SenticNet3.0 (Cambria, Olsher, and Rajagopal 2014), and SentiWordNet3.0 (Baccianella, Esuli, and Sebastiani 2010). For each event triple, we assigned an affective polarity based on the polarities of its component words. For an event triple e, we computed a polarity score using the following formula:

$$s(e) = \frac{1}{n}\Sigma_{w_i}^{n} lex\_score(w_i)$$

where w_i is a word in the event triple, and $lex\_score(w_i)$ is the polarity score given by the lexicon[8]. We also look for the presence of negation, and multiply the score by -1 if negation

[8]We use a default value of 0 if a word is not in the lexicon.

is found.

Since MPQA provides discrete labels, we used the following numeric scores for each word: positive=1, negative=-1, and neutral=0. The other three lexicons provide numeric polarity scores for each word, so we experimented with different thresholds to use the lexicons more aggressively or conservatively. The SenticNet (SNet) and SentiWordNet (SWN) lexicons have scores ranging from -1 to +1, so we assigned polarity scores as: *negative* if $s(e) \in [-1, -\lambda)$, *positive* if $s(e) \in (+\lambda, +1]$, and *neutral* otherwise. ConnotationWordNet (CWN) has scores ranging from 0 to +1, so we assigned polarity scores as: *negative* if $s(e) \in [0, .5 - \lambda)$, *positive* if $s(e) \in (.5 + \lambda, +1]$, and *neutral* otherwise. We experimented with λ values ranging from 0 to .4 in increments of .1. Table 3 shows the results for $\lambda = 0$ and the lambda value producing the best precision on the positive class for each lexicon.

	POSITIVE		NEGATIVE		
	Prec	#Events	Prec	#Events	Acc
G^{EV}	90.4	83	81.2	377	82.8
SWN (λ=0)	35.9	195	81.8	231	57.2
SNet (λ=0)	33.5	236	85.8	204	55.7
CWN (λ=0)	31.2	258	95.1	162	53.0
MPQA	64.3	28	90.6	170	47.0
SNet (λ=.2)	57.8	109	78.7	80	36.1
CWN (λ=.3)	44.1	143	92.5	40	28.0
SWN (λ=.3)	46.4	28	90.0	70	27.8

Table 3: Evaluation of polarity labels assigned by label propagation with G^{EV} and four sentiment lexicons

Table 3 shows the results for each sentiment lexicon as well as label propagation with our Event Context Graph G^{EV}. We present the results for all 460 affective events produced by G^{EV} and show precision for events labeled as positive, precision for events labeled as negative, and the accuracy over all events. First, we see that G^{EV} produced many more negative events than positive events. However, its precision when labeling an event as positive was over 90%. In contrast, the sentiment lexicons produced low precision for positive events, ranging from 31% for the most prolific output (CWN with λ=0) to 64% for the most conservative output (MPQA). When labeling events as negative, the sentiment lexicons all achieved $\geq 78\%$ precision. However, even the most prolific lexicon for the negative class, SWN, identified only 231 negative events, so failed to recognize many of the 377 negative events discovered by G^{EV}.

Overall, our analysis reveals that many affective events are not recognized as having polarity by traditional sentiment analyzers. Furthermore the accuracy of polarity labels assigned to events by sentiment lexicons is extremely low for positive events. These results further illustrate the need to acquire knowledge of affective events, and show that label propagation with event context graphs is a promising first step in this direction.

Conclusion

This research studied the problem of learning affective events associated with daily life from personal blogs. We constructed an *Event Context Graph* containing event and sentence nodes, and used label propagation to spread affective evidence from positive and negative sentences to event nodes. We explored three graph constructions by incorporating *event-sentence* edges, *sentence-sentence* edges, and *event-event* edges. Experimental results showed that our label propagation systems learned many affective events with good accuracy. We also analyzed several sentiment lexicons, and found that many of the affective events learned by our system cannot be recognized by the sentiment lexicons. In future work, we plan to create richer event representations and to explore additional mechanisms for inferring affective polarity based on discourse contexts.

Acknowledgements

This material is based upon work supported by the National Science Foundation under grant IIS-1450527. The authors thank Ashequl Qadir for insightful discussions about this work.

References

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Cambria, E.; Fu, J.; Bisio, F.; and Poria, S. 2015. Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Cambria, E.; Olsher, D.; and Rajagopal, D. 2014. Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *Twenty-eighth AAAI conference on artificial intelligence*.

de Marneffe, M.-C.; MacCartney, B.; and Manning, C. D. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

Deng, L., and Wiebe, J. 2014. Sentiment propagation via implicature constraints. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Deng, L., and Wiebe, J. 2015. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Deng, L.; Choi, Y.; and Wiebe, J. 2013. Benefactive/malefactive event and writer attitude annotation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Deng, L.; Wiebe, J.; and Choi, Y. 2014. Joint inference and disambiguation of implicit sentiments via implicature constraints. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*.

Feng, S.; Kang, J. S.; Kuznetsova, P.; and Choi, Y. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford University.

Gordon, A. S., and Swanson, R. 2008. Storyupgrade: Finding stories in internet weblogs. In *Proceedings of the Second International Conference on Weblogs and Social Media*.

Goyal, A.; Riloff, E.; and Daumé III, H. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*.

Goyal, A.; Riloff, E.; and Daumé III, H. 2013. A Computational Model forf Plot Units. *Computational Intelligence* 29(3):466–488.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177. ACM.

Kang, J. S.; Feng, S.; Akoglu, L.; and Choi, Y. 2014. Connotationwordnet: Learning connotation over the word+ sense network. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Lehnert, W. G. 1981. Plot units and narrative summarization. *Cognitive Science* 5(4):293–331.

Li, J.; Ritter, A.; Cardie, C.; and Hovy, E. 2014. Major life event extraction from twitter based on congratulations/condolences speech acts. In *Proceedings of Empirical Methods in Natural Language Processing*.

Li, H.; Mukherjee, A.; Si, J.; and Liu, B. 2015. Extracting verb expressions implying negative opinions. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Mohammad, S. M., and Turney, P. D. 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*.

Mohammad, S. M.; Kiritchenko, S.; and Zhu, X. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (SEMSTAR13)*.

Nielsen, F. Å. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*.

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

Qadir, A., and Riloff, E. 2014. Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Rao, D., and Ravichandran, D. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*.

Riloff, E.; Qadir, A.; Surve, P.; De Silva, L.; Gilbert, N.; and Huang, R. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Russo, I.; Caselli, T.; and Strapparava, C. 2015. Semeval-2015 task 9: Clipeval implicit polarity of events. In *Proceedings of the 9th International Workshop on Semantic Evaluation*.

Stone, P. J., and Hunt, E. B. 1963. A computer approach to content analysis: studies using the general inquirer system. In *Proceedings of the Spring Joint Computer Conference*.

Theobald, M.; Siddharth, J.; and Paepcke, A. 2008. Spotsigs: robust and efficient near duplicate detection in large web collections. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 563–570. ACM.

Velikovich, L.; Blair-Goldensohn, S.; Hannan, K.; and McDonald, R. 2010. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Vu, H. T.; Neubig, G.; Sakti, S.; Toda, T.; and Nakamura, S. 2014. Acquiring a dictionary of emotion-provoking events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*.

Zhang, L., and Liu, B. 2011. Identifying noun product features that imply opinions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Zhu, X., and Ghahramani, Z. 2002. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University.