

**AFFECTIVE POLARITY RECOGNITION AND HUMAN
NEEDS CATEGORIZATION FOR AFFECTIVE EVENTS**

by
Haibo Ding

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science

School of Computing
The University of Utah
December 2018

Copyright © Haibo Ding 2018

All Rights Reserved

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of Haibo Ding
has been approved by the following supervisory committee members:

<u>Ellen Riloff</u> ,	Chair(s)	<u>September 19 2018</u> Date Approved
<u>Feifei Li</u> ,	Member	<u>August 06 2018</u> Date Approved
<u>Thomas Fletcher</u> ,	Member	<u>August 06 2018</u> Date Approved
<u>Vivek Srikumar</u> ,	Member	<u>August 06 2018</u> Date Approved
<u>Saif Mohammad</u> ,	Member	<u>August 06 2018</u> Date Approved

by Ross Whitaker , Chair/Dean of
the Department/College/School of Computing
and by David B. Kieda , Dean of The Graduate School.

ABSTRACT

Though many improvements have been achieved in Natural Language Processing, the task of enabling computers to understand events that we experience is still far from achieved. Events that are stereotypically desirable (positive) or undesirable (negative) for experiencers are called *Affective Events*. Acquiring knowledge about affective events holds promise for obtaining a deeper understanding of narrative texts such as stories and conversations. In this dissertation, I present research on automatically learning knowledge about affective events: (1) the affective polarity of events indicates how experiencers are affected by the events (e.g., “I broke my leg” is typically undesirable), and (2) the human needs associated with affective events provide a general explanation for why an event is positive or negative (e.g., “I broke my leg” is negative because it violates the human need to be physically healthy).

My research designed two graph-based semi-supervised models to identify the affective polarity of events. The first *Event Context Graph* model identifies affective events by using discourse context and event collocation information. The second *Semantic Consistency Graph* model recognizes the affective polarity of events by optimizing the semantic consistency among events. Experimental results show that the Semantic Consistency Graph model outperformed previous methods and learned over 110,000 affective events with $>90\%$ precision for positive events and $>80\%$ precision for negative events.

My research also studied a new task of categorizing affective events into human need categories: *Physiological, Health, Leisure, Social, Financial, Cognition, and Freedom Needs*, which were developed to explain why events are affective. To automatically recognize human needs of affective events, I designed a co-training model that learns from unlabeled data by simultaneously training classifiers based on an event expression view and an event context view in an iterative learning process. Experimental results demonstrate that the co-training model achieved good performance, and outperformed each individual supervised classifier and a self-training model.

For my parents and my wife.

CONTENTS

ABSTRACT	iii
LIST OF FIGURES	viii
LIST OF TABLES	ix
ACKNOWLEDGEMENTS	xi
CHAPTERS	
1. INTRODUCTION	1
1.1 Affective Knowledge of Events	3
1.1.1 Affective Events	4
1.1.2 Affective Polarity of Events	4
1.1.3 Human Needs Implied by Affective Events	5
1.2 Semi-Supervised Methods for Acquiring Affective Knowledge of Events ...	6
1.2.1 Semi-Supervised Graph-based Models for Identifying Affective Polarity	7
1.2.2 A Co-Training Model for Recognizing Human Needs	9
1.3 Dissertation Claims and Research Contributions	11
1.4 Guide of This Dissertation	12
2. RELATED WORK	14
2.1 Knowledge Bases in NLP	14
2.2 Affect and Sentiment Analysis	15
2.2.1 The Concept of Affect	15
2.2.2 Sentiment Analysis	17
2.2.3 Implicit Sentiment Analysis	22
2.2.4 Sentiment Lexicons	23
2.3 Affective Event Analysis	25
2.4 Human Needs and Goals	28
2.4.1 Human Needs in Psychology	28
2.4.2 Goals and Desires for Text Understanding	29
2.5 Semi-Supervised Learning	30
2.5.1 Graph-based Semi-Supervised Learning	31
2.5.2 Self-Training and Co-Training	31
2.6 Chapter Summary	32
3. DATA SET AND EVENT REPRESENTATION	34
3.1 Personal Blogs Corpus	34
3.2 Event Frame Representation and Extraction	36
3.2.1 Basic Event Frame	36
3.2.2 Enhanced Event Frame	37

3.3	Gold Standard Polarity Annotations	41
3.3.1	Affective Polarity Annotation Guidelines	41
3.3.2	Manual Annotation Study	42
3.4	Chapter Summary	44
4.	EXTRACTING AFFECTIVE EVENTS FROM BLOGS WITH EVENT CONTEXT GRAPH MODEL	46
4.1	Motivation	47
4.2	Semi-Supervised Learning of Affective Polarity with Event Context Graph Model	48
4.2.1	Overview	48
4.2.2	Sentiment Sentence Classifier	48
4.2.3	Event Context Graphs	50
4.2.4	Variants of the Event Context Graph Model	52
4.2.5	Semi-Supervised Label Propagation	53
4.3	Evaluation	54
4.3.1	Baseline Systems	55
4.3.2	Evaluation Data Set and Metrics	56
4.3.3	Experimental Results	57
4.4	Analysis	58
4.4.1	Analysis of Learned Affective Events	58
4.4.2	Performance of Sentiment Lexicons	60
4.4.3	Evaluation on Randomly Sampled Events	62
4.4.4	Discussion of Limitations	63
4.5	Chapter Summary	65
5.	RECOGNIZING AFFECTIVE EVENTS USING SEMANTIC CONSISTENCY GRAPH MODEL	67
5.1	Motivation	68
5.2	Semantic Consistency Graph Model	69
5.2.1	Overview	69
5.2.2	Constructing the Semantic Relations Graph	70
5.2.3	Learning by Optimizing Semantic Consistency	72
5.2.3.1	Initialization	72
5.2.3.2	Semantic Consistency Metrics	73
5.2.3.3	Weight Normalization	74
5.2.3.4	The Objective and Update Functions	75
5.2.3.5	Improved Component Initialization	76
5.3	Evaluation	78
5.3.1	Performance of Affective Lexicons and Learning Models	78
5.3.2	Performance of the Semantic Consistency Graph Model	81
5.4	Analysis	82
5.4.1	Error Analysis	82
5.4.2	Quality and Quantity of the Learned Affective Events	84
5.5	Chapter Summary	85
6.	HUMAN NEEDS CATEGORIZATION OF AFFECTIVE EVENTS USING LABELED AND UNLABELED DATA	87

6.1	Human Need Categories and Annotations	89
6.1.1	Human Need Categories	89
6.1.2	Gold Human Need Annotations and Analysis	97
6.2	Categorizing Human Needs with Labeled and Unlabeled Data	100
6.2.1	Supervised Classification Models	101
6.2.1.1	Event Expression Classifiers	102
6.2.1.2	Event Context Classifiers	103
6.2.2	Semi-Supervised Models	104
6.2.2.1	Self-Training the Event Expression Classifier	104
6.2.2.2	Co-Training with Event Expression and Event Context Classifiers	105
6.3	Evaluation	107
6.3.1	Evaluation Metrics	108
6.3.2	LIWC Lexicon Baseline	108
6.3.3	Performance of the Event Expression Classifiers	109
6.3.4	Performance of the Event Context Classifiers	111
6.3.5	Performance of Self-Training and Co-Training Models	111
6.3.6	Analysis	114
6.4	Chapter Summary	115
7.	CONCLUSIONS AND FUTURE WORK	118
7.1	Research Summary and Contributions	118
7.2	Future Directions	120
7.2.1	Jointly Learning Affective Polarity and Human Needs	121
7.2.2	Recognizing Affective Polarity and Human Needs of Events in Stories and Conversations	122
7.2.3	Building a Hierarchical Knowledge Base of Affective Events	123
7.3	Summary	126
APPENDICES		
A.	AFFECTIVE POLARITY ANNOTATION GUIDELINES	127
B.	EXAMPLE EVENTS WITH AFFECTIVE POLARITY ANNOTATIONS	131
C.	DERIVATION FOR THE SEMANTIC CONSISTENCY GRAPH MODEL	134
D.	EXAMPLES OF AUTOMATICALLY LEARNED AFFECTIVE EVENTS	137
E.	HUMAN NEEDS ANNOTATION GUIDELINES	139
	REFERENCES	145

LIST OF FIGURES

3.1	Dependency Relations for Basic Event Frames	37
3.2	Polarity Annotation Confusions between Each Pair of Annotators.	44
4.1	Illustration of an Event Context Graph with Three Types of Edges	51
5.1	Semantic Relations Graph	69
6.1	Confusions between Two Annotators on Assigning Human Need Labels.	100
6.2	The Co-Training Model for Human Needs Categorization	106
6.3	Learning Curves of Self-Training and Co-Training Using the FoldAvg Metric . .	112
6.4	Learning Curves of Self-Training and Co-Training Using the TestAvg Metric . .	113
6.5	Confusions between Predictions and Gold Human Need Annotations.	115

LIST OF TABLES

3.1	Examples of Gold Standard Affective Events	43
3.2	The Distribution of Affective Events in the Gold Standard Data	44
4.1	Accuracy for the Top-Ranked Affective Events.	58
4.2	Top 50 Positive and 50 Negative Affective Events Produced with Label Propagation with G^{EV} , \emptyset Denotes Empty Element. Verbs that Usually Occur with a Particle Are Denoted with * (e.g., <i>screw up</i> , <i>black out</i> , <i>shut down</i>) to Help Readers Interpret the Likely Intended Phrase.	59
4.3	Evaluation of Polarity Labels Assigned by Label Propagation with G^{EV} and Four Sentiment Lexicons	61
4.4	Performance of ECG Model on Randomly Sampled Events	63
5.1	F1 Scores for Lexicons, Event Expression Classifiers, and Contextual Models . .	79
5.2	Precision and Recall for Lexicons, Event Expression Classifiers, and Contextual Models	80
5.3	Results for Semantic Consistency Graph (SCG) Model	82
5.4	Polarity Changes between Combo and SCG models	83
5.5	Precision and Recall Breakdowns for Combo and SCG Model	83
5.6	Correct and Incorrect Examples	84
5.7	Quality and Size of Affective Event Collections Extracted with Different Thresholds	85
6.1	Relations between the Human Need Categories Assigned to Affective Events for this Research and Maslow’s Hierarchy of Needs (MHN) and Fundamental Human Needs (FHN). The * Denotes that the MHN or FHN Categories Have Overlap with My Category in the Same Row, and the + Denotes that the MHN or FHN Categories Are a Subset of My Corresponding Category.	90
6.2	Affective Event Examples with Human Needs Category Labels.	98
6.3	Distribution of Human Need Categories (each cell shows the frequency and percentage).	98
6.4	Distribution of Affective Polarities under Different Human Need Categories . .	99
6.5	LIWC Mapping to Human Need Categories.	109
6.6	Performance of LIWC Baseline and Event Expression Classifiers	110
6.7	Performance of Event Context Classifiers	111
6.8	Performance of Self-Training and Co-Training	113

6.9	Breakdown of Results across Human Need Categories. Each Cell Shows Precision, Recall, and F1.	114
A.1	Event Examples and Their Corresponding Sentences.	128
B.1	Examples of Positive Events	131
B.2	Examples of Negative Events	132
B.3	Examples of Neutral Events	133
D.1	Examples of Automatically Learned Positive Events with Confidence 0.5	137
D.2	Examples of Automatically Learned Negative Events with Confidence 0.5	138

ACKNOWLEDGEMENTS

I am extremely grateful to my advisor, Professor Ellen Riloff; without her guidance, I would not have been able to complete this dissertation. Throughout my Ph.D. study, she has not only taught me many research skills such as how to find important research problems and solve them, but, more importantly, also demonstrated to me how to be a good professional researcher, and I learned a lot of research principles from her.

I would like to thank all of my committee members, Professor Thomas Fletcher, Professor Feifei Li, Dr. Saif M. Mohammad, and Professor Vivek Srikumar, for supervising my dissertation research. I thank Professor Vivek Srikumar for his suggestions and comments on my research work. I thank Dr. Saif M. Mohammad for detailed suggestions on the last part of my research work. I am also grateful to my MS thesis advisor, Professor Jingbo Zhu, for introducing me to the field of Natural Language Processing, and Dr. Muhua Zhu and Professor Tong Xiao for teaching me research skills and encouraging me to study abroad. I have been very lucky to have an opportunity to work as a summer intern at Robert Bosch Research, and I am grateful to Dr. Yifan He for mentoring my internship project. I had a wonderful summer at Bosch.

During my Ph.D. study in the NLP research group at the University of Utah, I have been so lucky to have a good number of friends and colleagues, with a special mention to Nathan Gilbert, Ruihong Huang, Youngjun Kim, Ashequl Qadir, Lalindra de Silva, Xingyuan Pan, Jie Cao, Tao Li, Tianyu Jiang, Yichu Zhou, Yuan Zhuang, Annie Cherkaev, Maks Cegielski-Johnson, Chi Zhang, Chengxu Ding, Mengyang Wang, Yan Zheng, Yang Gao, Fei Luo, Limou Wang, Xin Yu, and Xiaowan Li. I will always remember the numerous chats and the amazing time I have had with them. In addition, I am extremely grateful to Ashequl Qadir, Tianyu Jiang, Yichu Zhou, Xiaowan Li, and Yuan Zhuang for participating in my annotation study. Without their effort, my dissertation research would be incomplete.

I am grateful to my parents, Liangkai and Shenghua, my wife, Yuanyuan, and my

brother, Haixiang, for their unconditional support in my life. Without them, I would not be where I stand today. I am also thankful to my other family members who have supported me along the way.

Finally, last but not the least, I would like to acknowledge the funding sources that supported my Ph.D. research, which were the National Science Foundation under Grant Number IIS-1450527 and IIS-1619394. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

CHAPTER 1

INTRODUCTION

Current Natural Language Processing (NLP) systems can obtain much better performance than before in various tasks such as syntax analysis (e.g., part-of-speech tagging, parsing) and semantic analysis (e.g., named entity recognition, semantic role labeling), etc. However, the goal to let computers fully understand human language is still far from being accomplished. One major obstacle is that current NLP systems do not possess world knowledge like humans. For example, to understand a conversation or a story, we usually need to have world knowledge about events that impact people. As an illustration, if a friend said that “I got married, but I’m not happy”, people would probably ask “why?” because they have the world knowledge that if someone experienced the event “got married”, the person would typically feel happy. However, that friend was not. This contradiction makes people wonder about the reason behind it. As can be seen, acquiring world knowledge of events is a critical step to achieve such understanding.

Most of the events that we experience are mundane daily routines (e.g., “wake up”, “go to school/work”). However, some events will impact us in positive or negative ways. For example, we will be pleased when experiencing events like “having a campfire” or “having a birthday party”. On the other side, we will often be impacted negatively when experiencing events such as “dog passed away” or “broke a leg”. This dissertation will refer to events that are stereotypically positive or negative for experiencers as **Affective Events**. This dissertation presents approaches to learn two types of knowledge about affective events. The first type of knowledge is the affective polarity of events, which indicates whether an event is desirable (positive), undesirable (negative), or neutral. The second type of knowledge is the reason for events being affective, which explains why an event is affective (i.e., positive or negative).

Acquired knowledge of affective events can potentially benefit many NLP applications.

It will not only be useful to obtain better fine-grained sentiment/opinion analysis, but more importantly it will help achieve deeper understanding of narratives and conversations. For example, based on the recent study on sarcasm detection (Riloff et al., 2013), many sarcastic tweets are formed with a positive sentiment toward a negative situation. The learned knowledge of affective events can potentially improve the performance of sarcasm detection by recognizing negative situations. The acquired knowledge can also potentially help to produce plot unit representations (Goyal et al., 2010, 2013), which are knowledge structures to represent narrative stories (Lehnert, 1981), by recognizing affective states arising from positive or negative events. Affective knowledge can further be used to predict an experiencer's plans and goals (Schank and Abelson, 1977). To illustrate, given an event description "John broke his leg", we can understand that John is affected negatively and the reason is that he has a health problem. Then, we can further reasonably infer that John may have a goal to search for medical help. In addition, the knowledge can possibly benefit dialogue response generation (Li et al., 2016). With this knowledge, the dialogue agent can provide rational corresponding responses in a dialogue by correctly estimating a speaker's affective state given only event descriptions. For instance, knowing a speaker has experienced a very bad event, the dialogue agent should offer consolation. Conversely, the dialogue agent should offer congratulations if a speaker experienced some exciting and happy events. As a further example of a potential application, a mental health therapy system can benefit from understanding why someone is in a negative affective state. For example, if the triggering event for depression is "I broke my leg", then the reason is related to the person's health. However, if the triggering event is "I broke up with my girlfriend", then the reason is based on the person's social relations.

However, acquiring knowledge of affective events is challenging due to the following reasons. First, though learning the affective polarity of events is closely related to sentiment analysis, most existing sentiment analysis tools and resources fail to recognize many affective events that are implicitly affective. As demonstrated in our work (Ding and Riloff, 2016), many affective events are not recognized by one of the best sentiment classifiers (Mohammad et al., 2013), because they appear to be objective, factual expressions, and contain no explicit sentiment indicators. For example, affective events such as "I had a campfire" and "I dropped my phone into a toilet" are factual descriptions that cannot

be recognized by most existing sentiment analysis tools because these expressions do not explicitly express any sentiment.

Second, to the best of my knowledge, there is no prior research on the task of recognizing the reasons for events being affective. Answering the question of why an event is positive or negative sounds both trivial and difficult at the same time. For example, everyone knows the event “mom hugged me” is positive and “I woke up at 3am” is negative. They also understand what the events are about. However, if asked to explicate the reasons, there could be an infinite set of different answers. For example, the answers for why “I woke up at 3am” is negative could be “I’m sleepy”, “I don’t like waking up”, or “I need to rest”, etc. To give a general explanation using a small set of psychological categories, the research in this dissertation proposed to use the concept of “human needs” to explain why an event is positive (or negative), and to categorize affective events into one of the human need categories. This research also aims to design methods to achieve good performance on categorizing affective events with a small set of manually annotated data.

The rest of this chapter will first briefly introduce the two types of knowledge of affective events, and then present the dissertation claims and research contributions. The navigation of this dissertation will be presented at the end of this chapter.

1.1 Affective Knowledge of Events

As researchers are achieving better performance and understanding on various NLP tasks, they are increasingly recognizing the importance of world knowledge to natural language understanding. For example, it has been shown that world knowledge plays a critical role in answering elementary science questions (Li and Clark, 2015). Various knowledge bases have been created such as WordNet (Miller and Fellbaum, 1998), ConceptNet (Liu and Singh, 2004), and DBpedia (Lehmann et al., 2015), etc. These knowledge bases have proven to be useful in varied NLP tasks such as Question Answering (Li and Clark, 2015), Entity Linking (Mendes et al., 2011), and generating Abstract Meaning Representations (AMR) (Flanigan et al., 2016), etc. The research in this dissertation aims to learn world knowledge of affective events, which has not received much attention in previous research. This section will first present the concept of affective events, and then

briefly introduce two types of knowledge of affective events.

1.1.1 Affective Events

The term “**event**” in this dissertation refers to both **dynamic events** (e.g., “I ran a marathon”) and **static states** (e.g., “I am a student”), which we experience in our daily lives (e.g., “I woke up”, “I went to school”, and “we had a party”). Some events are beneficial and impact people positively such as “I got a new job”, while others are detrimental and impact people negatively like “I broke up with my girlfriend”. In this research, **affective events** are defined to be events that are stereotypically desirable (positive) or undesirable (negative) for experiencers, and they usually correspond to desirable or undesirable world states. For example, the event “have a party” is stereotypically desirable because it describes a recreation situation in which most experiencers would have a positive world state (e.g., having fun or feeling happy). When people experience events like “got a job”, “broke a record”, and “went to Disneyland”, they are usually impacted positively and have positive world states. However, it is possible that someone may have negative feelings when the person experiences the event “have a party” in real life. That said, it is a stereotypical and reasonable assumption that most experiencers would have a positive state unless contrary information is specifically given. Conversely, if someone experienced events such as “broke legs”, “was hit by a car”, and “failed the math exam”, the person will usually be affected negatively.

1.1.2 Affective Polarity of Events

The first type of event knowledge that this research aims to learn is the knowledge of how an event affects experiencers. Based on how people are affected, three categories (i.e., positive, negative, and neutral) are proposed to categorize events, which are referred to as the **affective polarity** of events. The definitions and examples for each affective polarity class are shown below.

- **Positive** events are those that most people would consider to be desirable, enjoyable, pleasant, or beneficial, etc. People are generally pleased if the events happen to them. Examples: “I danced with my friend”, “my confidence rose”, “I attended a show”.

- **Negative** events are those that most people would consider to be undesirable, unpleasant, or detrimental, etc. People are generally displeased if the events happen to them.

Examples: “my dog passed away”, “girl laughs at me”, “I was hit by a car”.

- **Neutral** (not affective) events are those that most people would not consider to be desirable or undesirable.

Examples: “I read a sentence”, “I opened the door”, “I packed up my bag”.

In this dissertation, the affective polarity of an event denotes the **prior polarity** of the event, which is independent of context, unless other information is given. As mentioned earlier, learning the knowledge of affective polarity (i.e., the prior polarity) can benefit many other NLP tasks such as fine-grained sentiment analysis, sarcasm detection, and story understanding.

1.1.3 Human Needs Implied by Affective Events

When we comprehend events, we not only know their affective polarities but also understand the reason why they impact us positively or negatively. For instance, when we comprehend the events “John’s house is on fire” and “John broke his leg” we do not only recognize that John is negatively affected but also the reasons why John is affected in that way. For the first event, John is affected negatively because John’s property was damaged. However, the second event is negative for John because John has a health problem. The explanation for why an event is affective (positive or negative) can potentially help NLP systems to achieve better understanding of events. Therefore, the second type of event knowledge that this research aims to learn is the knowledge of the reason for events being affective, i.e., why an event is positive or negative.

This research hypothesizes that the polarity of affective events can often be attributed to a relatively small set of **Human Need** categories. This is motivated by theories in psychology that explain people’s motivations, desires, and overall well-being in terms of categories associated with basic human needs, such as Maslow’s Hierarchy of Needs (Maslow et al., 1970) and Fundamental Human Needs (Max-Neef et al., 1991). Drawing upon these works, this dissertation proposes that the affective polarity of events often arises from 7 categories of human needs: *Physiological Needs*, *Health Needs*, *Leisure Needs*,

Social Needs, Financial Needs, Cognition Needs, and Freedom Needs. For example, “John broke his leg” has a negative polarity because it negatively impacts John’s Health. “John broke up with his girlfriend” is negative because it indicates a Social relation problem.

In this research, human needs are proposed to explain the reasons for events being affective because it is observed that whether an event is positive or negative for an experiencer is often highly correlated with whether the experiencer’s human needs are satisfied or not. As an illustration, the event “John broke his leg” is negative because John’s need to be physically healthy (Health Needs) is not satisfied. Conversely, the event “John went to Disneyland” is positive because John’s need to have fun (Leisure Needs) is satisfied. In daily life, especially in written or spoken language, people usually treat this information as common sense knowledge, which they don’t explicitly write out or speak, but they all understand. For example, it’s obvious that the event “John went bankrupt” is negative and people usually do not need to ask why it’s negative. However, asking why this event is negative can reveal that John’s need to be financially stable is not satisfied. Therefore, it is important to understand the reason for events being affective because it can provide deeper understanding of events by knowing whether experiencers’ human needs are satisfied or not. As illustrated earlier, understanding the reasons for events being affective can potentially benefit many NLP applications such as story understanding and mental health therapy systems.

1.2 Semi-Supervised Methods for Acquiring Affective Knowledge of Events

When children start to acquire knowledge about the world, we usually do not explicitly teach them everything about the world. Instead, we only show them a very small amount of world knowledge such as “this is an apple”, “that is a banana”, “you can not eat the toy car”, and “this event is bad for you”, etc. Then, they will learn additional knowledge by exploring the unknown world. In machine learning, a class of learning techniques that makes use of a small amount of labeled data (which is the knowledge being taught) and a large amount of unlabeled data (the unknown world) is called *semi-supervised learning*. In linguistics, researchers have found that children acquire language using bootstrapping mechanisms (Höhle, 2009), which is a kind of semi-supervised learning technique in ma-

chine learning. Inspired by this, this research aims to design semi-supervised machine learning methods¹ to acquire knowledge about affective events.

The practical motivation to use semi-supervised methods rather than supervised methods is that obtaining a large amount of manual annotations is expensive in terms of time and money. In addition, supervised models usually cannot be easily adapted to new tasks. For example, we may spend a lot of time and money to obtain a large amount of manual annotations in English when we want to learn the knowledge of events in English. However, supervised models trained on this data cannot be easily applied to learn knowledge of events in other languages (e.g., Arabic and Chinese). In contrast, semi-supervised learning has the advantage that models can be easily adapted to new tasks because they only require a small amount of manually acquired training data. In this research, two types of semi-supervised approaches were developed to learn knowledge of events. First, two graph-based semi-supervised methods were designed to identify affective polarity. Second, a co-training method was proposed to recognize the reasons (i.e., human need categories) to explain why events are affective.

1.2.1 Semi-Supervised Graph-based Models for Identifying Affective Polarity

One goal of this research is to identify affective polarity of events by designing semi-supervised methods. One intuition is to first use some events and contexts, for which polarities can be confidently recognized using existing sentiment analysis tools as initial supervision. Then, we can predict the polarity of other events based on the relations between the initial events and other events, and their contexts. Based on this intuition, this research designed two semi-supervised graph-based models to identify event polarity. The following paragraphs will briefly introduce these two models.

The first model is called the *Event Context Graph (ECG)* model, which identifies affective events using discourse contexts and event collocation statistics. For this research, events were extracted from a *personal blogs corpus*, which consists of blog stories about people's daily life and were extracted from a large set of Web blog posts. In the corpus, bloggers

¹Please note that some of the methods designed in this research work are actually weakly supervised methods, which are also called semi-supervised methods in this dissertation.

often express their sentiments or emotions toward exciting events such as a vacation or graduation, or unpleasant events such as an injury or job loss. The explicit sentiments sometimes are expressed directly about an affective event mentioned in the same sentence. For example, “I’m so happy that I’m graduating”. In many other cases, explicit sentiments are expressed in nearby sentences (discourse contexts). In the following example,

“Last weekend was really fun! We had a campfire at the beach.”

the explicit sentiment is expressed in the first sentence, and the event is described in the following sentence. Based on this observation, this research explores the idea of harvesting affective events from a large collection of personal story blogs by identifying events that frequently occur in positive or negative discourse contexts. In addition, it is also observed that highly correlated events (that frequently co-occur in the same document) may have similar affective polarity. For instance, we would expect the events “girl was hurt” and “girl cried” to frequently co-occur in blog posts discussing accidents involving children. Based on these two intuitions, this dissertation describes a semi-supervised learning algorithm to extract affective events by propagating affective evidence from event contexts containing explicit sentiments to events, and then from some events to other highly correlated events. Specifically, the ECG model obtains noisy supervision by automatically identifying sentence contexts with strong polarity values with an existing sentiment analysis classifier. Then, the model extracts affective events by iteratively spreading polarity evidence from sentence contexts to events using a Label Propagation algorithm (Zhu and Ghahramani, 2002).

Though the ECG model learns some affective events with good precision, it is difficult to extend the algorithm to a much larger data set because of the large number of sentence nodes in the model. In addition, the previous algorithm does not exploit semantic relations among events. Therefore, this dissertation also introduces a *Semantic Consistency Graph (SCG)* model to induce a large set of affective events more efficiently by relying on semantic consistency and only using noisy evidence from existing sentiment analysis tools and resources as supervision. This method is motivated by three intuitions, which are formulated as three types of semantic consistencies. The first type of semantic consistency is that semantically similar events will usually have similar affective polarities.

For example, given the two similar events “have party” and “have celebration”, they have similar meanings and they are both positive. The second type of semantic consistency is that semantically opposite events will usually have opposite affective polarities. For instance, the events “I won” and “I did not win” have opposite semantic meanings, and their affective polarities are also opposite (i.e., “I won” is positive and “I did not win” is negative). The third type of semantic consistency is that the affective polarity of an event often originates from its components. For example, if “birthday party” is positive, this model hypothesizes that most events mentioning “birthday party” will be positive too. Based on these intuitions, the learning method seeks to achieve a state where the affective polarities of events are consistent with their semantic relations. Specifically, a graph was first built using events and their individual components as nodes, then a weakly supervised, graph-based propagation algorithm was designed to assign affective polarities by maximizing the semantic consistency (i.e., minimizing the semantic inconsistency) in the graph. One advantage of the SCG model is that it only uses noisy evidence obtained from existing sentiment analysis tools and resources as supervision, which makes it easily scalable to large data sets without any manual annotation effort.

To summarize, the first ECG model builds a graph using discourse contexts and collocation statistics, and then propagates polarity evidence from event contexts with automatically assigned polarities to other unlabeled events. The second SCG model builds a graph using semantic relations among events and their components, and assigns initial polarities (which are noisy) using existing sentiment resources. Then, it infers the true polarities of events by optimizing the semantic consistency in the graph. The key differences between these two models are that (1) the graphs in the two models represent different types of event information, and (2) the polarity information is estimated differently, i.e., the polarities are propagated based on affective contexts and collocations in the ECG model, while the polarities are inferred by optimizing semantic consistency with respect to event representations in the SCG model.

1.2.2 A Co-Training Model for Recognizing Human Needs

The models in the previous section aim to identify affective polarity of events, i.e., whether an event is positive, negative, or neutral. This section will briefly introduce

another semi-supervised method to recognize the reasons for events being affective, i.e., why an event is positive or negative. This dissertation formalizes this problem as a task to categorize affective events into a small set of human need categories: *Physiological Needs*, *Health Needs*, *Leisure Needs*, *Social Needs*, *Financial Needs*, *Cognition Needs*, and *Freedom Needs*. The goal is to assign the most appropriate human need category label to each affective event based on the stereotypical (default) meaning of the event, which is independent of any specific context.

Observing that the human need category of an event can not only be recognized using the event expression itself but also the collection of contexts surrounding the event, this dissertation proposed a semi-supervised co-training approach to tackle this problem. First, the most obvious approach is to train a supervised classifier using the words in event expressions as features for recognizing the human need categories. For instance, we can train a logistic regression classifier using the bag-of-words features from event expressions (e.g., bag-of-words features {I, was, hit, by, a, car} for the event “I was hit by a car”). In this dissertation, the classifier trained using features from event expressions is called an *Event Expression Classifier*. In addition, events were originally extracted from a large collection of personal story blogs that contain many instances of the events in different sentences. Therefore, the contexts surrounding instances of an event can also provide strong clues about the human need category associated with the event. As an illustration, given an event “I broke my leg”, there are many sentences in the blog data set mentioning this event such as “I broke my leg when I was hit by a big guy in my last year of football” and “While studying in LA I had a motorcycle accident, and I broke my leg”. This example shows that the contexts of event mentions are also strong indicators for recognizing an event’s human need category. Based on this idea, this dissertation built an *Event Context Classifier* using features from the contexts surrounding event mentions.

Since event expression classifiers and event context classifiers represent two different views of an event, one hypothesis is that these two views can potentially be complementary to each other, and unlabeled events confidently selected by one classifier can be valuable training instances to benefit the another classifier in an iterative learning process. Based on this idea, this dissertation designed a co-training model for human needs categorization by taking advantage of both an event expression classifier and an event context

classifier to achieve better performance using only a small amount of labeled events and a large amount of unlabeled events.

1.3 Dissertation Claims and Research Contributions

The primary claims and contributions of this dissertation are as follows:

Claim #1. Many affective events in personal stories can be identified and assigned prior polarities using graph-based semi-supervised learning.

Knowledge of affective events plays an important role for understanding narrative stories and conversations. However, there are no existing resources of affective events. In this research, two graph-based semi-supervised models were designed to identify affective events and assign polarities to them. Experimental results demonstrate that the two models can identify many affective events and assign polarities to them with good precision, and achieving better performance than methods based on previous techniques and resources.

The first Event Context Graph (ECG) model builds a graph using discourse and collocation information, and spreads the polarity evidence from seeding sentence nodes to other unlabeled nodes. It has been demonstrated that the discourse contexts and collocation statistics can increasingly improve the performance of identifying affective polarities of events. The second Semantic Consistency Graph (SCG) model creates a graph using three types of semantic relations: semantic similarity, semantic opposition, and event component relations. Affective polarities of events are estimated by optimizing the semantic consistency in the graph. Experimental results show that the three types of semantic relations can increasingly improve the accuracy of recognizing affective polarity. Further analysis shows that the SCG model performs better than the ECG model, and learns over 110,000 affective events with >90% precision for positive events and >80% precision for negative events. This demonstrates that many affective events can be identified and assigned polarities using graph-based semi-supervised learning.

Claim #2. Affective events can be automatically categorized into a small set of human need categories by co-training models with views based on event expressions and event contexts.

Analyzing the reasons for events being affective has not been explored in prior research. A contribution of this research is formalizing this problem as a multiclass clas-

sification task that assigns a human need category to each affective event. This research demonstrated that most affective events can be categorized into a small set of 7 human need categories through a manual annotation study in which human annotators achieved good annotation agreements on this task.

To automatically recognize human need categories, two supervised classifiers were built: an event expression classifier that uses features from event expressions alone, and an event context classifier that uses surrounding context features of event mentions. Experimental results demonstrate that both of these classifiers can achieve relatively good performance using only a small amount of labeled data. This suggests that both the event expression view and the event context view can be sufficient to decide the human need category of an event. Based on the idea that these two views can be complementary to each other, a semi-supervised co-training model was developed to effectively combine the two classifiers and exploit unlabeled data. Experimental results show that the co-training model achieved better performance than each individual classifier, which indicates that the two individual classifiers can benefit each other. This research demonstrates that affective events can be categorized into the 7 human need categories by co-training models with an event expression view and an event context view.

1.4 Guide of This Dissertation

The rest of this dissertation is organized as follows:

Chapter 2 reviews the concept of affect, and presents related work on sentiment and emotion analysis, and prior work on affective event analysis. This chapter also presents related work on human needs and goals, and several semi-supervised learning algorithms used in this dissertation.

Chapter 3 presents the story corpus used in this research, and introduces the event frames used to represent events. This chapter also describes methods for extracting event frames from the story corpus.

Chapter 4 describes the details of the Event Context Graph model designed to acquire affective events. This chapter first describes the method for constructing an event context graph using event discourse contexts and event collocations, and the label propagation algorithm used to learn affective events. At the end, this chapter evaluates the event

context graph model and compares it with methods based on previous resources.

Chapter 5 describes the details of the Semantic Consistency Graph model. Different from the Event Context Graph model, this model estimates affective polarity by optimizing the consistency of semantic relations in a graph instead of propagating the polarity from context nodes to event nodes. This chapter first describes the method used to build the graph with three types of semantic relations. Then, this chapter presents an iterative learning algorithm for identifying affective polarity. Finally, this chapter evaluates the semantic consistency graph model and compares it with previous methods.

Chapter 6 first introduces definitions of the 7 human need categories, and presents a manual annotation study for obtaining gold human need category labels for affective events. Then, this chapter describes details of an event expression classifier, an event context classifier, and a co-training framework used to categorize affective events. At the end, this chapter presents experiments for evaluating the proposed methods using gold annotations.

Chapter 7 summarizes the research work in this dissertation, and discusses potential research directions for future work.

CHAPTER 2

RELATED WORK

The goal of this research is to learn affective polarity and human needs knowledge about affective events. To give a review of related research, this chapter first introduces knowledge bases that have been created in prior research. Next, this chapter reviews definitions of the concept of “affect”, and presents related research on sentiment analysis, implicit sentiment analysis, and sentiment lexicon induction. Then, it describes prior research on affective event analysis, which is most closely related to the work in this dissertation. In addition, this chapter presents two theories on human needs in prior psychology research, and overviews prior research on “goal” analysis, which is related to the human need analysis studied in this dissertation. At the end, this chapter describes semi-supervised learning algorithms related to the research in this dissertation.

2.1 Knowledge Bases in NLP

Acquiring knowledge is an important step for natural language understanding, and various knowledge bases have been created. In the earlier days, researchers constructed knowledge bases manually. For example, WordNet (Miller and Fellbaum, 1998) is a manually created database of English words. Each word in WordNet is associated with senses and connected with other words through different semantic relations (e.g., synonym, hypernym, and hyponym). ConceptNet (Liu and Singh, 2004) is a knowledge base containing concepts and commonsense knowledge that was crowdsourced or created by experts. Recently, researchers developed a new version (ConceptNet 5) (Speer and Havasi, 2013) by automatically incorporating other available knowledge bases such as DBpedia (Lehmann et al., 2015) and Wikipedia. Cyc (Lenat and Guha, 1993) is another manually created knowledge base that contains basic concepts and rules about how the world works.

Building a knowledge base manually is time consuming and expensive. In recent work, researchers have designed automatic methods to acquire facts and commonsense

knowledge from the Web such as Freebase (Bollacker et al., 2008), WebChild (Tandon et al., 2014), Open IE (Fader et al., 2011; Mausam et al., 2012), NELL (Mitchell et al., 2015), YAGO (Rebele et al., 2016), and Probase (Wu et al., 2012). These knowledge bases usually contain entities, semantic concepts, and relations among entities and concepts. For example, “cat” and “animal” are two concepts, and the relation between them is “isA” (i.e., “cat isA animal”). Most of these automatically built knowledge bases are much larger than those that were manually created. For instance, YAGO contains 17 million entities (Rebele et al., 2016), while ConceptNet 2.0, which is manually built, contains 300 thousand concepts (Liu and Singh, 2004). In addition, these knowledge bases have been successfully used in NLP applications such as Question Answering (Li and Clark, 2015), Entity Linking (Mendes et al., 2011), and producing Abstract Meaning Representations (AMR) (Flanigan et al., 2016).

Besides these knowledge bases of facts, entities, semantic concepts, and semantic relations, there have been many research works focusing on acquiring sentiment lexicons, which will be discussed in Section 2.2.4. This dissertation aims to acquire affective polarity and human needs knowledge about affective events, and create a new type of knowledge resource for the NLP community.

2.2 Affect and Sentiment Analysis

2.2.1 The Concept of Affect

This section reviews definitions of “affect” in both computer science and other disciplines. Some researchers use “affect” as an umbrella term that covers several concepts. For example, Fleckenstein (1991) suggested to use “affect” as an umbrella term to include emotions, feelings, moods, preferences, beliefs, attitudes, motivations, and evaluations. Scherer (2000) proposed to categorize affective states into five classes: emotions, moods, interpersonal stances, attitudes, and personality traits. In addition, following the tradition in *affective computing* (Picard, 1995), Jurafsky and Martin (2016) used “affective” to mean emotion, sentiment, personality, mood, and attitude.

Some other researchers, especially people in psychology and cognition, defined the term “affect” differently. Tomkins (1962, 1963) defined “affect” to be a biological portion of emotion. In a later research, (Nathanson, 1994, 58) gave a further explanation about

Tomkins' definition and described "affects" as "groups of hard-wired, preprogrammed, genetically transmitted mechanisms that exist in each of us". Tomkins' definition and Nathanson's explanation both suggest that "affect" corresponds to a biological mechanism that is built in us. When there is a stimulus, this mechanism will be triggered and a series of biological events will happen, and when we are aware of this, we would often use the word "feeling" to indicate that we are affected. In recent works, researchers in psychology tend to have similar definitions about "affect", and try to give definitions for several "affect"-related concepts. For example, Shouse (2005) distinguished feeling, emotion, and affect as three concepts. He explained that feeling was "a sensation that has been checked against previous experiences and labeled"; emotion was "the projection or display of a feeling", which could be "either genuine or feigned"; and affect was "a non-conscious experience of intensity", which was also "a moment of unformed and unstructured potential". Recently, Munezero et al. (2014) provided more subtle differences between five closely related concepts. They explicated that feelings were "person-centered, conscious phenomena"; emotions were "preconscious social expressions of feelings and affect influenced by culture"; affect was "a predecessor to feelings and emotions"; sentiments were "partly social constructs of emotions that develop over time and are enduring"; opinions were "personal interpretations of information that may or may not be emotionally charged", which means that opinions may not be aligned with sentiments in some cases. These definitions suggest that affect happens before feeling; when people become aware of the affect, then it becomes people's feelings or emotions. Following the definitions in the prior research (Munezero et al., 2014; Shouse, 2005), the research in this dissertation treats "affect" as "a non-conscious experience of intensity" and "a predecessor to feelings and emotions", and focuses on studying affective events that are stereotypically desirable or undesirable for experiencers based on world knowledge.

The affective polarity of events, which is independent of context, is prior knowledge based on the most stereotypical understanding of events. For example, the event "we had a party" is stereotypically desirable because it describes a recreation situation in which most experiencers would typically have fun and feel happy. As mentioned in Chapter 1, events that are stereotypically desirable (positive) or undesirable (negative) for experiencers are called *Affective Events*. In my research, affective events usually correspond to desirable

or undesirable word states, which is the most typical understanding without considering specific contexts. When someone experiences a positive (or negative) event, the person will typically have a positive (or negative) world state, but this is not absolutely certain because in the real world a person's world state can also depend on other factors. For example, the typical understanding of the event "I got laid off" is that the speaker lost his/her job and felt bad, so the event is negative and the speaker typically had a negative world state. However, this is the most stereotypical understanding; in the real world, the speaker may have a positive world state in some specific cases. For example, the speaker may hate the job because the payment is not good and the work is boring, and the speaker has already found a new high-paying job at the time of being laid off. In this case, the speaker may have a positive world state even though the typical understanding of the event is negative.

2.2.2 Sentiment Analysis

Learning affective polarity of events is generally related to sentiment analysis, which is a popular research topic in NLP and many other fields (e.g., Machine Learning and Data Mining). This section gives a broad view of prior research on sentiment analysis.

Given different contexts, the term "sentiment analysis" may refer to different things. For example, in a survey (Pang and Lee, 2008), "sentiment analysis" is broadly defined as "the computational treatment of opinion, sentiment, and subjectivity in text". "Sentiment analysis" and "opinion mining" are used interchangeably, and both denote the same field of study (Pang and Lee, 2008). In a recent book (Liu, 2012), "sentiment analysis" is also called "opinion mining" and is defined as "the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes". Others may refer to "sentiment analysis" as a specific task. For instance, a natural language processing textbook (Jurafsky and Martin, 2016) describes "sentiment analysis" as a text categorization task that aims to extract "the positive or negative orientation that a writer expresses toward some object". Most of the prior research on sentiment analysis has focused on recognizing sentiments or opinions expressed in text. However, this field is being expanded increasingly. For example, many new research problems have been

proposed such as research on studying implicit sentiment including opinion implicatures (Deng and Wiebe, 2014; Deng et al., 2014), connotation of words (Kang et al., 2014), and affective events (Ding and Riloff, 2016, 2018b). Sections 2.2.3 and 2.3 will give more details about these recent research works on implicit sentiment analysis.

Subjectivity analysis is also closely related to sentiment analysis. Subjectivity analysis mainly focuses on determining whether a given text is subjective or objective (Wiebe et al., 2001; Wilson et al., 2005; Esuli and Sebastiani, 2006; Pang and Lee, 2008). In prior research (Wiebe et al., 2001), subjectivity in natural language was defined to refer to “aspects of language used to express opinions and evaluations” (Banfield, 1982), and subjective sentences “present private states of characters - states of an experiencer holding an attitude, optionally toward an object” (Wiebe, 1994), while the primary intention of an objective sentence is to “objectively present material that is factual to the reporter” (Bruce and Wiebe, 1999). For example, the sentence “At several different levels, it’s a fascinating tale.” is subjective, while the sentence “Bell Industries Inc. increased its quarterly to 10 cents from seven cents a share” is objective. As explained in another prior research (Esuli and Sebastiani, 2006), objective text is usually factual, describes “a situation or event without expressing a positive or a negative opinion”. This explanation suggests that objective language expressions are not associated with positive or negative sentiments/opinions, but this does not mean that all subjective texts are positive or negative even though they are often associated with positive/negative sentiments. As shown in prior research (Wilson et al., 2005), subjective expressions can also be neutral. For example, the sentence “John says the hospital feels no different than a hospital in the states” is subjective but does not express any positive or negative sentiments.

Research in sentiment analysis has often focused on studying overall sentiment expressed in different levels of text (e.g., document-level or sentence-level). The document-level task aims to analyze overall sentiment expressed in a whole document. For example, Pang et al. (2002) applied three machine learning methods (Naive Bayes, Maximum Entropy, Support Vector Machines) to determine whether an author expressed an overall positive or negative sentiment in a review. Turney (2002) designed unsupervised methods to classify product reviews as *recommend* or *not recommend*. The sentence-level task aims to recognize sentiment expressed in a single sentence or phrase. For instance, Yu

and Hatzivassiloglou (2003) studied both supervised Naive Bayes classifiers and unsupervised methods to distinguish subjective sentences from objective sentences and determine whether the opinions expressed in subjective sentences are positive or negative. Wilson et al. (2005) studied *contextual polarities* of phrases that depend on contexts where the phrases are mentioned. In their work, Wilson et al. (2005) built classifiers to first classify a phrase as polar (i.e., positive or negative) or neutral, then determine contextual polarity values of polar phrases. Recently, there have been many research works using deep learning techniques for analyzing the sentiment of sentences such as recursive neural networks (Socher et al., 2013), and convolutional neural networks (Kim, 2014).

Opinion mining is another concept that is closely related to sentiment analysis. In research focused on extracting evaluations or judgments towards products from reviews (Dave et al., 2003; Pang et al., 2002), opinion mining and sentiment analysis have been used interchangeably and can refer to the same task. For example, in reviews, consumers may express a positive/negative sentiment or opinion towards a specific product. However, in some other cases, an “opinion” does not necessarily mean a sentiment. Opinions can be beliefs, views, and thoughts without expressing any sentiments. For example, given the sentence “I believe there are two cats in the room”, we know that the speaker expressed an opinion but not a positive or negative sentiment. In opinion analysis, there has been research focusing on identifying or extracting opinion holders (or sources) and targets. An opinion holder or source is an entity that expresses an opinion, and an opinion target is an entity or proposition at which the opinion is directed (Wiegand et al., 2016). To recognize opinion holders and targets, various methods and features have been proposed such as Conditional Random Fields and lexico-syntactic patterns (Choi et al., 2005), syntactic structures (Kessler and Nicolov, 2009), and semantic roles (Kim and Hovy, 2006; Johansson and Moschitti, 2013). Recently, Wiegand et al. (2016) also studied the problem of extracting opinion holders and targets on opinion compounds.

In addition, some researchers have studied other aspects of opinions. For example, researchers have studied the views that an opinion expression evokes (Wiegand and Ruppenhofer, 2015; Wiegand et al., 2016). Wiegand and Ruppenhofer (2015) proposed to study three different views for verbal predicates (i.e., the agent view, the patient view, and the speaker view). Later, Wiegand et al. (2016) merged the agent view and the patient view into

one single actor view. The *actor views* are “expressions conveying sentiment of the entities participating in the event denoted by the opinion word”. For example, in the sentence “All representatives praised^{actor} the final agreement”, the sentiment conveyed by the word “praised” is the actor’s view, while the *speaker views* are “expressions conveying sentiment of the speaker of the utterance”. For example, in the sentence “Sarah excelled^{speaker} in virtually every subject”, the sentiment conveyed by the word “excelled” is the speaker’s view. Some other researchers have tried to generate abstractive summaries for opinionated texts (e.g., reviews for products, arguments for a controversial social issue) (Hu and Liu, 2004; Lerman et al., 2009; Wang and Ling, 2016).

Besides the research works discussed above, many other tasks on sentiment analysis have also been proposed. Some research focused on analyzing sentiments or opinions toward various aspects of products (Hu and Liu, 2004; Jiang et al., 2011). For example, Titov and McDonald (2008) studied sentiment expressed toward food, environment, and service aspects of restaurants. Instead of classifying sentiment into two categories (i.e., positive and negative), researchers have also studied fine-grained sentiment categories. For instance, Nakov et al. (2016) categorized sentiment expressed in tweets into five categories, i.e., highly-positive, positive, neutral, negative, and highly-negative. These fine-grained sentiment categories can provide a general understanding about the strength of sentiments. Recently, more research has been proposed to study the degrees of positivity or negativity (i.e., sentiment intensity) associated with words or phrases (Kiritchenko et al., 2016; Kiritchenko and Mohammad, 2017). Kiritchenko and Mohammad (2016) also analyzed the effect of different modifiers (e.g., negators, modals, and degree adverbs) on the sentiment of words being modified. Some researchers designed methods to recognize stance in tweets and analyzed the interactions between stance and sentiment (Sobhani et al., 2016; Mohammad et al., 2016, 2017).

Besides the research on sentiment, researchers have also studied the task of detecting emotions in text. Most of the research on detecting emotions tries to classify a given text into a distinct set of emotion categories such as Anger, Joy, Sadness, and Fear, which were proposed in prior research (Ekman, 1992). For example, researchers have built emotion classifiers to classify Web blogs and tweets into emotion categories (Yang et al., 2007; Roberts et al., 2012). Researchers have also tried to associate words and phrases asso-

ciated with emotion categories via crowdsourcing (Mohammad and Turney, 2010, 2013). Qadir and Riloff (2014) designed automatic methods to learn emotion indicators such as hashtags, hashtag patterns, and phrases from tweets. Recently, researchers have begun to study emotion intensity (i.e., the strength of emotion) associated with a given text. For example, Mohammad (2018) manually created an affect intensity lexicon with real-valued scores of intensity for four basic emotions using the best-worst scaling annotation method. Researchers have also proposed a task of detecting intensity of emotion in tweets and provided annotated data with intensity values (Mohammad and Bravo-Marquez, 2017; Mohammad and Kiritchenko, 2018b). Additionally, there are some research works studying emotions evoked by art (Mohammad and Kiritchenko, 2018a), and emotionality of metaphorical expressions (Mohammad et al., 2016).

In addition, sentiment analysis has been widely used in various applications (Pang and Lee, 2008; Liu, 2015). For example, some researchers used sentiment analysis to forecast revenues of movies (Sadikov et al., 2009; Asur and Huberman, 2010; Joshi et al., 2010), and to predict one-day-ahead stock index by analyzing sentiments in tweets (Si et al., 2013). In addition, researchers have proposed methods to predict political election results by analyzing public opinions (O'Connor et al., 2010; Bermingham and Smeaton, 2011; Chung and Mustafaraj, 2011; Sang and Bos, 2012).

The research presented in this dissertation aims to identify affective events from personal stories and understand the reasons for events being affective. The first part of my research, i.e., the task of recognizing affective events, is closely related to prior research on sentiment analysis, but has difference emphases. First, this dissertation focuses on events. For the purpose of my research, each event is represented with a predicate and three event arguments (i.e., agent, theme, and prepositional phrase), which are heuristically extracted using dependency relation rules. Second, most prior sentiment analysis tasks have focused on predicting the sentiment in subjective language expressions. Though evaluation data sets used in prior research may contain objective expressions, often these objective expressions have been annotated as neutral. However, many affective events studied in this dissertation are factual, objective expressions. Many methods have been proposed for sentiment analysis including supervised methods (Jurafsky and Martin, 2016; Mohammad et al., 2013; Pang et al., 2002; Socher et al., 2013) and unsupervised methods

(Turney, 2002; Yu and Hatzivassiloglou, 2003). The models designed in this dissertation for recognizing affective events are related to semi-supervised graph-based methods that have been previously used for sentiment lexicon induction (Feng et al., 2013; Rao and Ravichandran, 2009; Velikovich et al., 2010). A key difference is that the novel graph structures used in this dissertation were designed based on event collocations, discourse information, and semantic relations. In addition, I also designed a new iterative learning method for the Semantic Consistency Graph model to recognize affective polarity of events.

2.2.3 Implicit Sentiment Analysis

Most research mentioned in the previous sections focused on explicit sentiment, which is usually directly expressed in documents or sentences, although some sentiment lexicons may also contain some words or phrases that do not explicitly express sentiment but are highly associated with positive or negative sentiment based on their lexical semantics (e.g., “celebration”, “catastrophe”). There is a growing interest in recognizing sentiment that is indirectly or implicitly expressed. The term “*implicit sentiment*” was used in previous research (Deng and Wiebe, 2014; Deng et al., 2014) to mean sentiment that is not explicitly expressed. Deng and Wiebe (2014) found that opinion toward entities or events may not be explicitly expressed, and proposed to use opinion implicature rules to infer opinions toward different entities. For example, in the sentence “The bill would lower health care costs, which would be a tremendous positive change across the entire health-care system.”; though the writer only explicitly expressed a positive opinion toward the idea of lowering health care costs, we can infer that the writer is negative toward the “health care costs” based on that she/he is positive toward the event “lowering”, which is bad for the “health care costs”. Some researchers have focused on studying implicit sentiment conveyed by individual words or phrases. For example, Zhang and Liu (2011) noticed that some objective nouns and noun phrases do not indicate sentiment in general domains. However, in some specific domains, these nouns or phrases would imply sentiment. For example, “valley” implies negative sentiments in the sentence “Within a month, a valley has formed in the middle of the mattress” (Zhang and Liu, 2011). Feng et al. (2013) and Kang et al. (2014) also found that some objective words often indicated connotative and nuanced sentiment. For example, the words “intelligence” and “Ph.D.” usually have positive connotations while

“cancer” and “war” often have negative connotations. Though the denotative or surface meanings of these words are objective, they often implicitly express the sentiment of the writer. For example, in the description “corruption is a cancer to our society”, the writer implicitly expressed a negative sentiment toward corruption through the word “cancer”. They proposed graph-based methods to automatically learn connotations of words and word senses. In addition, Rashkin et al. (2016) claimed that writers can subtly imply several types of connotations (e.g., writer’s perspective, entities’ perspectives, effect on entities, entity values, and entities’ mental states) through the choice of a predicate, and designed *connotation frames* to represent and organize these various types of connotations. For example, given a predicate “violate”, if someone writes a sentence “#Agent violates #Theme”, then we can use connotation frames to infer that the writer has negative perspective toward #Agent but a positive perspective toward #Theme, #Agent has a negative perspective toward #Theme, and #Theme has a negative perspective toward #Agent. The predicate has a negative effect on #Theme, and #Theme is in a negative state. In addition to the research on implicit sentiment analysis, there have been many research efforts studying the affect of events, which will be discussed in Section 2.3.

The research in this dissertation is closely related to implicit sentiment analysis because many affective events are often objective and factual expressions (e.g., “we went to Disneyland”, “my phone stopped working”), but impact people positively or negatively.

2.2.4 Sentiment Lexicons

Most of the research mentioned in the previous sections aims to recognize sentiment in context. However, researchers have found that knowing the *prior polarity* (Wilson et al., 2005) of words or phrases, which is independent of context, can improve sentiment analysis. *The prior polarity* of a word or phrase denotes the polarity determined by the meaning of the word or phrase without considering the context. Various research has been conducted to construct resources by associating prior polarity to words or phrases. The resulting resources are often called *sentiment lexicons*. It has been shown that sentiment lexicons are useful for many tasks. For example, Mohammad et al. (2013) built a state-of-the-art tweet sentiment analysis system by using sentiment lexicons as key features.

Because of their usefulness, many sentiment lexicons have been created manually or

automatically. For example, General Inquirer (Stone et al., 1968) is a manually built lexicon containing both positive and negative words. LIWC (Pennebaker et al., 2007) is also a manually created lexicon containing words associated with sentiment polarities and “psychologically meaningful” lexical categories. The MPQA Subjectivity Lexicon (Wilson et al., 2005) includes both words from the General Inquirer and words that were automatically learned and manually reviewed (Riloff and Wiebe, 2003). SenticNet (Cambria et al., 2014, 2015) is a sentiment lexicon compiled from knowledge bases like DBPedia (Bizer et al., 2009). SentiWordNet (Baccianella et al., 2010) is an automatically learned lexicon that contains synsets from WordNet (Miller and Fellbaum, 1998). ConnotationWordNet (Feng et al., 2013; Kang et al., 2014) is an automatically learned lexicon containing both explicit sentiment words (e.g., happy, good) and objective words (e.g., promotion, cancer) that convey connotative sentiment. In addition, Hu and Liu (2004) designed an automatic method to extract a sentiment lexicon from product reviews. Choi and Wiebe (2014) induced a sense-level lexicon in which each verb sense from WordNet (Miller and Fellbaum, 1998) was automatically labeled with +/- effect polarity.

Besides these lexicons associated with positive or negative polarity labels, there are some lexicons that contain words or phrases that are associated with emotion categories. For example, the NRC Emotion Lexicon (Mohammad and Turney, 2010, 2013) contains a list of words associated with eight basic emotions. Qadir and Riloff (2014) designed a weakly supervised method to induce lists of hashtags, hashtag patterns, and phrases associated with five emotions. Recently, some researchers have also created lexicons of words and phrases associated with sentiment intensity (i.e., the degree of positivity or negativity) (Kiritchenko et al., 2016; Kiritchenko and Mohammad, 2017), and real-valued scores of emotion intensity (i.e., the strength of emotion) (Mohammad, 2018).

Instead of building a lexicon of individual words or phrases, this dissertation focuses on recognizing affective events, which are represented as frame-like structures with 4 components. Learning the affective polarity of events is challenging because the polarity value of an event is not simply a combination of the polarities of the words in the event frame. As demonstrated in this dissertation, using existing sentiment lexicons to predict the polarities of events is not sufficient, so methods need to be designed to learn the affective polarity of events are needed.

2.3 Affective Event Analysis

Events are an important component of narrative stories. Therefore, to comprehend narrative stories, we need to understand the events in them. There has been some research focusing on learning the affective polarities and opinions about events. For example, Goyal et al. (2013) tackled the problem of automatically generating plot units, which are conceptual knowledge structures for representing narrative stories (Lehnert, 1981). In their corpus of AESOP fables, they discovered that 36% of +/- affect states originated from good or bad situations (i.e., positive or negative events). To recognize these good or bad situations, they designed two semi-supervised methods (Goyal et al., 2010, 2013) to identify verbs that impart positive or negative polarity on their patients (i.e., Theme semantic roles of verbs). The first method was based on the co-occurrence between verbs and evil or kind agents. The second method used the Basilisk bootstrapping algorithm (Thelen and Riloff, 2002) with conjunction contextual patterns. The resulting verbs, called *Patient Polarity Verbs* (PPVs), were then used to recognize good or bad states for their patients. Rashkin et al. (2016) extended this idea by studying the effect that verbs have on both their agents and themes. They also analyzed the writer's perspective and entities' perspective toward both the agent and theme of a given predicate, and acquired a lexicon of verbs associated with connotation labels. Their research of the effect that a verb has on its agent or theme is closely related to the research presented in this dissertation. In contrast to their research, which analyzes the effect of a verb based on only the given verb, this dissertation aims to understand how experiencers are affected by an event that consists of not only a verb but also 3 arguments. For example, in their research, the verb "eat" has a positive effect on its agent. However, the affective polarity of an event is determined by the meaning of the whole event representation rather than just the verb (e.g., the event "I ate poison" is negative for the speaker).

Deng et al. (2013) created a corpus with annotations of events that positively or negatively affect entities. They called these events benefactive (goodfor) or malefactive (badfor) events, which were later renamed as +/-effect events (Deng and Wiebe, 2014). They observed that sometimes opinions may not be expressed directly but can be inferred using +/-effect events, and investigated how +/-effect events can be used to improve the recognition of writer's opinions toward other entities. They designed a Loopy Belief

Propagation model to propagate opinions among entities using opinion implicature rules based on +/-effect events. Later, Deng et al. (2014) developed an unsupervised optimization framework (Integer Linear Programming) to incorporate implicature rules as constraints, and achieved better performance on both opinion detection and +/-effect event identification tasks. In addition, Deng and Wiebe (2015b) extended the MPQA corpus (Wiebe et al., 2005) by adding annotations on entities and events. The resulting MPQA 3.0 can be used to develop and evaluate systems for entity and event-level opinion analysis. Deng and Wiebe (2015a) also proposed probabilistic soft logic models to improve the performance of entity/event-level opinion analysis by incorporating explicit sentiments, inference rules, and +/-effect event information. This research mainly focused on identifying positive/negative opinions toward entities and events in specific contexts. In contrast, this dissertation aims to acquire general prior knowledge about stereotypically positive or negative events for the person experiencing the event, which is independent of contexts. In addition, their research mainly focused on analyzing opinions toward entities or events in texts, which are usually subjective expressions. However, many of events studied in this dissertation are factual and objective, which do not express any opinions. For example, the event “We went to Disneyland” is typically desirable and the event “I woke up at 3am” is undesirable for experiencers, but the speakers do not express any opinions in these expressions.

Recent work on “major life event” extraction (Li et al., 2014) collected major life events from tweets. In their work, a bootstrapping approach was first initialized with replies representing two types of speech acts: congratulations and condolences. Then, a large collection of tweets were clustered using topic models and manually filtered and labeled. A set of classifiers was trained to label tweets with respect to 42 event categories corresponding to different topics. Finally, for each type of event, they extracted event properties (e.g., names of spouse for wedding events) from the labeled tweets. Since the two types of speech acts (congratulations and condolences) usually highly correlate with positive and negative events, many of the extracted events are affective events even though their work did not aim to recognize the affective polarities of the events.

Besides the research of learning positive (good) or negative (bad) events, researchers also studied events that can provoke emotions. For example, Tokuhisa et al. (2008) tackled

a Japanese emotion classification task using a large collection of emotion-provoking events that were extracted from the Web using a lexical pattern “someone be #EMOTION #that #EVENT”. In the pattern, #EMOTION is an emotion word, #that is a connective word (e.g., “that”), and #EVENT is a subordinate clause, which will be extracted as an emotion provoking event. For example, given the sentence “I was disappointed that it suddenly started raining”, the pattern will extract an event “it suddenly started raining” that provokes the disappointment emotion. In their experiments, Tokuhisa et al. (2008) showed that the learned events were useful for predicting the emotions of a speaker conversing with a dialog system. Recently, Vu et al. (2014) extended this idea by designing bootstrapped learning approaches to learn emotion-provoking events. Their learning method takes a similar lexical pattern (i.e., “I am #EMOTION that #EVENT”) and several seed words for six emotions as input, then it iteratively expands patterns and acquires emotion-provoking events at the same time. They demonstrated that their method can extract emotion-provoking events better than the method that only uses the pattern (Tokuhisa et al., 2008). However, in the blog story used in this research, affective events that are associated with this type of lexical pattern are very rare. As they concluded in their work, better methods need to be explored to improve the coverage of extracting emotion provoking events.

Recognizing affective events has been demonstrated to be useful in many NLP applications. For example, Riloff et al. (2013) observed that recognizing affective events can help to detect sarcastic tweets. Their work showed that sarcasm often arises from the juxtaposition of a positive sentiment with a negative situation (event). Therefore, they developed a semi-supervised bootstrapping algorithm to automatically learn explicit sentiment expressions and phrases corresponding to negative situations, and showed that the learned negative situations can help to recognize sarcastic tweets. Recently, Reed et al. (2017) learned a set of lexico-functional linguistic patterns, which were extracted using the AutoSlog-TS system (Riloff, 1996), and their corresponding affect polarities. Their research demonstrated that the learned patterns can be helpful in predicting first-person affective reactions to daily life events. SemEval-2015 Task 9 (Russo et al., 2015) is a shared task that aims to determine whether the event mentioned in a given sentence is *pleasant* or *unpleasant*. However, their annotations of event polarity are based on specific instances of

an event in context.

In addition, there has been some recent work studying affective polarity in figurative language. For example, Qadir et al. (2015) explored methods to automatically recognize affective polarity in similes, which are special forms of figurative language that compare two unlike things. For example, the expression “Jane swims like a dolphin” expresses a positive sentiment toward Jane’s swimming ability. Though a simile can also describe an activity or a state (e.g., “he runs like a cheetah”), most of the events studied in this dissertation are not similes.

2.4 Human Needs and Goals

Besides affective polarity, the research in this dissertation also aims to learn human needs implied by affective events, i.e., to understand the reason why an affective event is positive or negative. This section overviews research on related topics.

2.4.1 Human Needs in Psychology

To the best of our knowledge, there is little work on analyzing human needs in NLP. However, psychological theories have been proposed to explain people’s motivations, desires, and overall well-being in terms of categories associated with basic human needs, such as Maslow’s Hierarchy of Needs (Maslow et al., 1970) and Fundamental Human Needs (Max-Neef et al., 1991). Maslow’s Hierarchy of Needs was developed to understand human behavior and motivations. The hierarchy of needs is often depicted using a pyramid with the most basic needs at the bottom and higher level needs at the top. The hierarchy implies that basic needs should be met before higher level needs, though Maslow acknowledged that different levels of needs could occur at any time. Maslow’s theory enables us to understand people’s motivation for their actions. Specifically, human needs in Maslow’s theory are categorized as *Physiological Needs*, *Safety Needs*, *Belonging and Love Needs*, *Esteem Needs*, and *Self-actualization Needs*. In Maslow’s later years, he added an additional level of *Self-transcendence Needs* into his theory (Maslow, 1971)

Another human needs theory developed by (Max-Neef et al., 1991) is called Fundamental Human Needs, which was developed to study community strength and weakness. In this theory, Max-Neef et al. (1991) proposed the following categories: *Subsistence Needs*,

Protection Needs, Affection Needs, Understanding Needs, Participation Needs, Leisure Needs, Creation Needs, Identity Needs, and Freedom Needs. Their theory claims that in contrast to Maslow's theory, there is no hierarchy; human needs are all interrelated and interactive. They also claimed that fundamental human needs are finite and classifiable, would not change over time, and are constant through all human cultures (Max-Neef et al., 1991).

In this dissertation, I developed a set of human need categories by selecting and separating some categories from these theories. The detailed definitions and examples of these new categories will be described in Chapter 6.

2.4.2 Goals and Desires for Text Understanding

Recognizing human needs is related to research on understanding goals and desires. The concept of "goals" was proposed to understand narrative stories by Schank and Abelson (1977) who stated that understanding can be achieved when a series of actions can be reasoned to fulfill goals. To better understand different types of goals in stories, they proposed to categorize goals into following classes: *satisfaction goal, enjoyment goal, achievement goal, preservation goal, crisis goal, instrumental goal, and delta goal* (Schank and Abelson, 1977). Some of these classes overlap with the human needs categories studied in this dissertation. For example, the satisfaction goal, which is defined as a recurring biological need, is closely related to the Physiological need category. The achievement goal, which is defined as the realization of "some valued acquisition or social position", is related to the Finance need and Social need studied in this dissertation. Some researchers found that analyzing the affective states of characters in stories is closely related to the satisfaction and violation of goals. In her plot unit work, Lehnert (1981) explained that achieving a goal successfully would result in a positive affect state, and otherwise, a negative affect state. Dyer (1983) also claimed that if a goal is achieved, then the characters may feel happy, joyous, or glad, while if a goal is thwarted or suspended, characters may feel unhappy, upset, or sad. In narrative stories, goals often are explicitly mentioned or could be inferred. In many cases, the concepts of goals and human needs can be used interchangeably. Comparing to goals, the subtle difference is that human needs usually are more fundamental. For example, John may have a goal to buy a hamburger and fries, but the fundamental reason is that John has the need (Physiological needs) to eat food.

Recently, some researchers tried to recognize people's wishes, goals, and desires in text. For example, Goldberg et al. (2009) analyzed the WISH corpus obtained from the Times Square Alliance, and proposed a task to classify individual sentences as being wishes or not. They also designed classifiers using various features (e.g., words, patterns) to detect wishes. Their experiments demonstrated that wish templates generated from the WISH corpus can improve wish detection performance on product reviews and political discussion posts. Chaturvedi et al. (2016) proposed a task of identifying if a desire expressed in a given text was fulfilled. They first created two corpora with desire fulfillment labels, then designed both unstructured and structured models to detect whether a desire was fulfilled or not. They also showed that the system, which models the evolution of a story as latent states, achieved the best performance. Inspired by this work, Rahimtoroghi et al. (2017) introduced a new data set (DesireDB) that contains labels for identifying desire expressions and whether a desire was fulfilled. As stated by Rahimtoroghi et al. (2017), narrative and sentence structures in DesireDB are more complex than the two data sets used in the prior work (Chaturvedi et al., 2016). To determine whether a desire was fulfilled or not, Rahimtoroghi et al. (2017) designed supervised models with various features (e.g., contextual and word embedding features).

2.5 Semi-Supervised Learning

Semi-supervised learning is a group of machine learning methods that aim to achieve good performance by taking advantage of a small amount of labeled data and a large amount of unlabeled data. Semi-supervised learning has gained a lot of research interests from different research fields such as Machine Learning, Natural Language Processing, and Data Mining, etc. Various semi-supervised learning algorithms have been proposed such as graph-based semi-supervised learning (Zhu, 2005), semi-supervised support vector machines (Bennett and Demiriz, 1999), co-training (Blum and Mitchell, 1998), and semi-supervised deep neural networks (Kingma et al., 2014). Since the research in this dissertation aims to design semi-supervised learning methods to learn knowledge of affective events, this section will overview some of the semi-supervised learning algorithms that are closely related to the methods used in this dissertation.

2.5.1 Graph-based Semi-Supervised Learning

Graph-based semi-supervised methods have been widely used in various applications, such as text classification (Talukdar and Crammer, 2009; Orbach and Crammer, 2012), part-of-speech tagging (Subramanya et al., 2010; Das and Petrov, 2011), and semantic parsing (Das and Smith, 2011). Some researchers also designed graph-based methods to acquire sentiment lexicons (Feng et al., 2013; Rao and Ravichandran, 2009; Velikovich et al., 2010).

The first part of my research designed two graph-based semi-supervised learning models to learn the affective polarity of events. The first Event Context Graph (ECG) model is motivated by the Label Propagation (LP) algorithm (Zhu and Ghahramani, 2002), which was proposed to solve classification problems. The LP algorithm iteratively propagates label information from a small set of labeled instances (seeding information) to other unlabeled instances through graph edges. The second Semantic Consistency Graph (SCG) model is inspired by the semi-supervised graph-based methods proposed by Zhou et al. (2003) and Subramanya and Bilmes (2011). Motivated by these methods, the SCG model uses noisy labels as initialization, and computes the inconsistency between the values on two connected nodes as the KL-divergence of the two values instead of the square loss measure. Comparing to previous methods, the objective function of the SCG model is more complex, which models not only similarity relations but also opposition relations between nodes.

2.5.2 Self-Training and Co-Training

The second part of the research in this dissertation designed self-training and co-training approaches to recognize human needs implied by affective events. Self-training is a semi-supervised learning method that obtains a better model by exploiting the predictions of a supervised classifier on unlabeled data. Self-training first trains a model using only labeled data, then iteratively re-trains the model using both the original training data and its own predictions on unlabeled data. It has been demonstrated that self-training can successfully improve the performance of many NLP tasks but often offers a precision/recall trade-off (i.e., recall is often increased but precision may be sacrificed). For example, self-training has been used to improve the performance of information extraction (Ding and Riloff, 2015) and syntactic parsing (McClosky et al., 2006).

Co-training (Blum and Mitchell, 1998) is another semi-supervised learning framework that trains models by exploiting multiple independent views of the data. It has been previously used in NLP tasks such as Word Sense Disambiguation (Mihalcea, 2004) and Coreference Resolution (Phillips and Riloff, 2002). Based on co-training, Kumar and Daumé III (2011) proposed an effective spectral clustering method that has access to multiple views of a data point. In addition, sentiment analysis has been improved by using co-training methods. For instance, Wan (2009) designed a better Chinese sentiment analysis system by using both Chinese features and English features (i.e., cross-lingual views). Recently, Xia et al. (2015) improved sentiment analysis performance by co-training a model with both the original and anonymous views. This dissertation designed a new co-training method to recognize human needs implied by affective events by exploiting both the event expression view and the contextual view of an event.

2.6 Chapter Summary

This chapter overviews prior works related to the research in this dissertation. First, this chapter reviewed knowledge bases that have been manually or automatically acquired in prior research. Most of these knowledge bases consist of facts, concepts, and semantic relations, which is different from the knowledge of affective polarity and human needs about affective events studied in this dissertation. Then, this chapter discussed the definitions of the concept “affect” in prior research and prior works on the analysis of sentiments, subjectivity, opinions, emotions, and implicit sentiments. It also reviewed sentiment lexicons that are commonly used in sentiment analysis. Most of these prior research focused on analyzing sentiments or opinions expressed in text, which is subjective. In contrast, this dissertation aims to understand how experiencers are stereotypically affected by events, most of which are factual and objective expressions and may not express any sentiments or opinions. Some of the prior research has studied affective events denoted by individual verbs. However, the affective polarity of an event is often determined by its verb (i.e., predicate) and arguments rather than just the verb. In this dissertation, I represent each event with a frame-like structure that consists of 4 components: Agent, Predicate, Theme, and Prepositional Phrase, and analyze the affective polarity of events based on the whole event representations.

Second, this chapter briefly introduced two psychological theories, based on which the human need categories were developed for categorizing affective events. This chapter also discussed prior research on understanding goals and desires, which are closely related to the research on categorizing affective events based on their human needs studied in this dissertation. In contrast to these prior research, this dissertation studied a new research problem of understanding the reasons for events being affective and claimed that the majority of affective events extracted from blog posts can be categorized into a small set of human need categories.

In addition, this chapter reviewed related research on graph-based semi-supervised learning, self-training, and co-training, which were used in this dissertation.

CHAPTER 3

DATA SET AND EVENT REPRESENTATION

This chapter presents the details of the text corpora used in this dissertation. Since this dissertation aims to learn knowledge of affective events about daily life, I collected a large of Web blog posts discussing about daily life stories. The detailed preprocessing steps will be described in this chapter. Since my research studies events independent of contexts, events are represented as tuples consisting of Agent, Predicate, Theme, and Prepositional Phrase, which are extracted using dependency relations. This chapter provides details about how events are represented and how they are collected from the text corpora. In addition, this chapter presents a manual annotation effort to study the prevalence of affective events and create a gold standard data with polarity labels for evaluation.

3.1 Personal Blogs Corpus

The text corpora used in this research are from ICWSM 2009 and 2011 data challenges provided by Spinn3r.com. The ICWSM 2009 Spinn3r data set (Burton et al., 2009) contains 44 million blog posts that originated between August 1st and October 1st, 2008. The ICWSM 2011 Spinn3r data set (Burton et al., 2011) consists of over 386 million texts, including blog posts (133 million), news articles (15 million), and other social media data (238 million), which originated between January 13th and February 14th, 2011. For the research in this dissertation, I used all of the ICWSM 2009 data set and the blog post portion of the ICWSM 2011 data set, which together contain over 177 million raw blog posts.

Since my research aims to study events in daily life stories and not all blog posts are personal stories, the raw blogs were filtered to identify personal story posts. First, the text contents were extracted from the original XML or HTML files. Observing that many of the texts came from Web domains such as craigslist.org and answers.yahoo.com and are not narrative stories, I only used the texts originating from blogging sites, which were manually identified. Though both the ICWSM 2009 data set and 2011 data set were

provided by Spinn3r, they are different because the 2011 data set contains a lot of blog posts containing pictures and several sentences or just titles describing the pictures (e.g., artlimited.net and tumblr.com). Therefore, I manually selected blog sites for each data set separately. For the ICWSM 2009 blog data, the sites are livejournal.com, wordpress.com, blogspot.com, spaces.live.com, typepad.com, and travelpod.com. For the ICWSM 2011 blog data, the sites are blogspot.com, myspace.com, livejournal.com, wordpress.com, travelpod.com, and wattpad.com. Second, personal story blog posts were identified using a personal story classifier (Gordon and Swanson, 2009), which was developed to identify stories that “are primarily a first person description of events in the life of the author” (Gordon and Swanson, 2009). Due to classification errors, some of the resulting blogs actually are not true narrative stories, so I used an additional heuristic rule to remove posts that contain no first-person mentions, i.e., if a story blog post does not have any first-person mention (e.g., I, me, we, and us), then it would be removed. Third, near-duplicate story blogs were also removed using SpotSigs (Theobald et al., 2008) because I noticed that some stories were re-posted many times by different bloggers. This process resulted in 1,383,425 personal story blogs that were further processed using NLP tools. Specifically, each story was first tokenized, split into sentences, and annotated with Part-Of-Speech and Named Entity labels using the Stanford CoreNLP tool (Manning et al., 2014), then was further parsed using the SyntaxNet (Andor et al., 2016) dependency parser. Specifically, I used the collapsed dependencies with propagation of conjunct dependencies, i.e., the `CCPropagatedDependencies` setting defined in the prior work (De Marneffe and Manning, 2008). Because the SyntaxNet parser does not produce collapsed dependencies directly, I used the Stanford CoreNLP tool to convert outputs of the SyntaxNet into collapsed dependencies. Please note that in the research presented in Chapter 4, I used Stanford parser to obtain dependency relations, but later I noticed that the SyntaxNet parser does better than Stanford parser, especially on sentences containing infinitive verb constructions. Therefore, I used the SyntaxNet parser to obtain dependency relations for the research presented in Chapters 5 and 6.

3.2 Event Frame Representation and Extraction

This section presents definitions and extraction methods for two event frame representations. The basic event frame was first proposed and used in the research presented in Chapter 4, which represents an event as a triple that contains an Agent, Predicate, and Theme. Later, I found that some important information of events cannot be represented using the basic event frame. Therefore, I designed an enhanced event frame to incorporate Preposition Phrase components. Both of these two event representations were extracted from the text corpora described in the previous section using dependency relations. The details of the extraction methods for each type of event frame are described below.

3.2.1 Basic Event Frame

I define basic event frames as triples consisting of 3 components: **Agent**, **Predicate**, and **Theme**. Each component is represented with a single word (one exception is that the predicate can also be attached with a negator), which is extracted based on dependency relations. First, I identify each verb as a potential predicate for an event. Since the negator modifier of a predicate is critical to determine the meaning of an event. For example, “I win” is semantically opposite to the event “I didn’t win”. Therefore, the negator modifier of an predicate is also extracted using the *neg* dependency relation and attached to the predicate, if present.

To extract the Agent and Theme component for an event, I used heuristic rules based on dependency relations rather than semantic role labeling. Specifically, I used two types of rules for active and voice sentences separately. For active voice verb phrase constructions, the heads of a verb’s subject (Agent) and its direct object (Theme) are extracted using the *nsubj* and *dobj* dependency relations. Extracted events are normalized by lemmatizing their subject, direct object, and verb. For example, the sentence (a) in Figure 3.1 would produce the event $\langle \text{they, have, party} \rangle$. For passive voice constructions, the heads of a verb’s subject (Theme) and its agent (Agent) are extracted with the *nsubjpass* and *agent* relations. Each component is also lemmatized. For example, Figure 3.1(b) shows a sentence in the passive voice, which produces the event $\langle \text{man, be.killed, by.police} \rangle$. To make it easily readable, I keep the verb in its past tense and append “be”, and append “by” to the agent.

In cases when an active voice construction does not have a direct object, or a passive

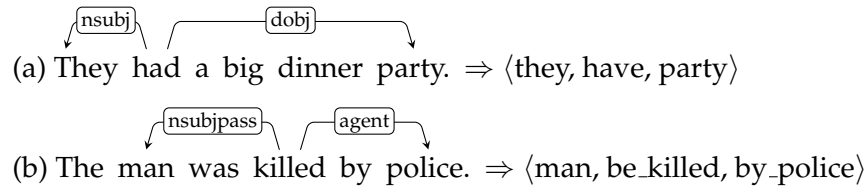


Figure 3.1: Dependency Relations for Basic Event Frames

voice construction does not have an agent, one of these elements may be absent. If both the Agent and Theme components of an event frame are absent, the event will only have a Predicate, which is too abstract and insufficient for people to understand the event. Therefore, I restrict that each basic event frame must have a Predicate and at least one other component (an Agent or Theme).

3.2.2 Enhanced Event Frame

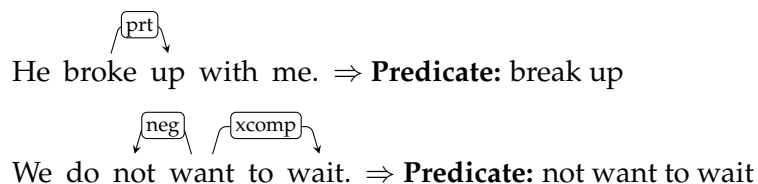
During the research in this dissertation, I first developed the basic event frame, but later I came to realize that the meanings of some events cannot be adequately represented using the basic event frame. Therefore, I subsequently designed an enhanced event frame to incorporate more information into an event representation.

Compared with the basic event frame, the enhanced event frame contains the following additional information. First, a prepositional phrase in an event is important to understand the meaning of the event. Given two event descriptions "I got into an argument" and "I got into the college", the basic event frame will only extract "I get", and loses the semantic information of the two prepositional phrases: "into a argument" and "into the college", which are essential to distinguish between dramatically different event types. Therefore, the enhanced event frame has an additional prepositional phrase component to capture the meaning of a prepositional phrase. Second, the basic event frame only extracts single words for Agent and Theme components, which is inadequate to capture their meanings. Because, for many events, their Agent and Theme components are often Named Entities and noun compounds that require more than one words to sufficiently represent the concept (e.g., "oil price", "White House"). Therefore, in the enhanced event frame, each of these components is composed with a **Minimal Noun Phrase**, which can also be a Named Entity or a noun compound. Specifically, a minimal noun phrase is

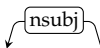
recognized in the order of Named Entities, nominals with noun premodifiers, pronouns, and single nouns. If a minimal noun phrase is modified by a possessive pronoun, then the possessive pronoun is also extracted using the *poss* relation and attached to the phrase. The possessive pronoun can be used to distinguish whether the event is related to the first person (i.e., the blogger). Third, I noticed that there are many stative expressions (most of them are predicate adjective constructions) that can also be useful to identify a person's affect state (e.g., "my dad is brave"). Therefore, I also allow the theme to be an adjective to extract stative events. Fourth, the predicate in the enhanced frame can also be associated with a particle and a infinitive verb or gerund besides a negator. For example, a predicate can be "not want to leave". Finally, using the enhanced event representation, active and passive event constructions are normalized. For example, "I was killed by him" and "he killed me" are both represented by the same frame: " \langle he, kill, me, - \rangle ".

Specifically, an enhanced event frame consists of 4 components and is represented as: \langle **Agent, Predicate, Theme, PP** \rangle , in which **PP** denotes the prepositional phrase component. In the enhanced event frame, each event argument is approximated using dependency relation rules rather than through semantic role labeling. The details of the extraction method for each event component are presented below.

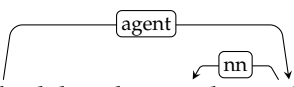
A **Predicate** is extracted for each verb, and can include a negator, a particle, and a infinitive verb or gerund, which are separately extracted using *neg*, *prt*, and *xcomp* dependency relations, if present. For the examples shown below, the Predicate of the first sentence is "break up", and the Predicate of the second one is "not want to wait".



An **Agent** is represented with a minimal noun phrase and extracted using the *nsubj* relation from an active voice sentence, and the *agent* relation from a passive voice sentence. Given the following two event descriptions, their Agents are "White House" and "stock market", respectively. The ORG denotes ORGANIZATION named entities.

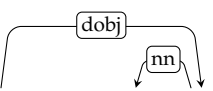


 White House rejected the proposal. ⇒ **Agent:** White House
 ORG ORG

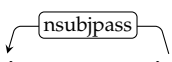


 He was shocked by the stock market. ⇒ **Agent:** stock market

A **Theme** can be a minimal noun phrase, and is extracted using the *dobj* or *nsubjpass* relations from an active or passive voice sentence correspondingly, if present. For example, the Themes of the following two sentences are “oil price” and “project”.




 He predicted the oil price. ⇒ **Theme:** oil price

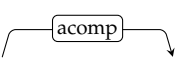


 The project was rejected by the government. ⇒ **Theme:** project

As mentioned earlier, a Theme can also be used loosely to allow an adjective to fill this component when it cannot be extracted using *dobj* or *nsubjpass* relations. When the predicate of an event is a copula, the Theme is extracted using the *cop* relation. For example, in the sentence below “my dad is brave”, the Theme is “brave”. Otherwise, the Theme is extracted using the *acomp* relation. For example, the Theme is “beautiful” in the sentence below “she looks beautiful”.



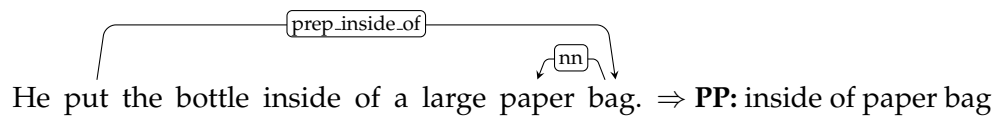
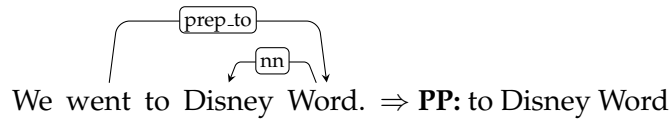
 My dad is brave. ⇒ **Theme:** brave



 She looks very beautiful. ⇒ **Theme:** beautiful

A **PP** component is extracted using the collapsed *prep_X* relation in which *X* denotes a preposition word (e.g., “in”, “on”) or phrase (e.g., “inside of”, “along with”). The preposition word or phrase (*X*) is extracted with the preposition object (a minimal noun phrase) using the *prep_X* relation. Only a single PP is extracted for each event to prevent the representation from becoming overly specific though multiple PPs are common and can be important. If multiple PPs are attached to a predicate, only the closest one to the predicate will be extracted. The following shows two concrete examples of how PP components are

extracted using dependency relations.



As described above, active and passive event constructions are normalized using the enhanced event representation. In addition, all words in event frames are also lemmatized for normalization. For some events, both Agent and Theme are missing, in which case, the meanings of these events are not accurately captured. Therefore, an event frame is required to must have an Agent or a Theme, but PP is optional. When the verbs are “be” and “have”, both Agent and Theme are required, which is used to filter event fragments such as “⟨I, be, -, -⟩” and “⟨we, have, -, -⟩”.

The research presented in Chapters 5 and 6, which uses the enhanced event representation, aims to analyze affective events from the perspective of the first person (i.e., the blogger). (Please note that the research in Chapter 4 does not analyze events from the first person perspective.) Therefore, to make sure that extracted events are first person related, only events that satisfy at least one of the following criteria are collected: (1) The event has a first-person reference such as a first-person pronoun (e.g., “I went to school”, “the sellers accepted our offer”). (2) The event mentions a family member (e.g., “mom”). This assumes that the affective state of the blogger usually parallels that of family members (e.g., “mom is sick” is undesirable for both mom and the blogger). A list of 92 family terms were manually compiled and used to identify family member mentions. (3) The event does not mention any other people. This heuristic assumes that the event pertains to the blogger (e.g., “the computer died”). An entity is identified as “other people” if it is a second- or third-person pronoun, a PERSON Named Entity, or nominal person mention based on WordNet (e.g., “plumber”). Specifically, I obtained normal person words by extracting all synonyms and hyponyms of the word “person” from WordNet. I did not extract events that only mention other people because they may be describing someone else’s experience, not the blogger’s. These rules were manually defined and could probably be improved

with discourse analysis in future work.

Using the enhanced event frame representation, I collected 19,794,187 unique event frames from the preprocessed personal blogs corpora described in Section 3.1. Events with low frequencies mean that the events are not commonly discussed in personal stories. In addition, I observed that low frequent events tend to be noisy. Because of using automatic preprocessing tools (e.g., POS and NER tagging, and dependency parsing) that are not perfect, there are some errors when extracting events from the automatically preprocessed blogs posts. Therefore, events with frequency less than 5 were filtered. Finally, this resulted a set of 571,424 unique event frames, which is called **AffectEvent** data set.

3.3 Gold Standard Polarity Annotations

Recognizing affective events can be useful for many other NLP tasks such as fine-grained sentiment analysis, sarcasm detection, and narrative story understanding. However, a key question for research on this topic is: how prevalent are affective events? To answer this question and to create a test set for evaluation, I conducted a manual annotation study to label a random set of events extracted from the personal blogs corpus with affective polarities. This section first presents annotation guidelines for affective polarity, and then shows details of a manual annotation study on affective events.

3.3.1 Affective Polarity Annotation Guidelines

For the research of learning affective polarity, I initially defined four affective polarity labels: POSITIVE, NEGATIVE, NEUTRAL, and MIXED for categorizing events. The definitions and examples of each label are described below, and the original annotation guidelines are presented in Appendix A.

POSITIVE: An event that most people would consider to be desirable, enjoyable, pleasant, or beneficial. People are generally pleased if the event happens to them.

Examples:

⟨I, dance, -, with my friend⟩ ⟨I, have, birthday party, -⟩
 ⟨I, be, -, in Disney World⟩ ⟨I, find, job, -⟩

NEGATIVE: An event that most people would consider to be undesirable, unpleasant, or detrimental. People are generally displeased if the event happens to them.

Examples:

⟨dog, pass away, -, -⟩ ⟨girl, laugh, -, at me⟩
 ⟨car, injure, mom, -⟩ ⟨I, lose, job, -⟩

NEUTRAL: An event that most people would not consider to be positive or negative.

Examples:

⟨I, sit, -, -⟩ ⟨I, open, door, -⟩
 ⟨I, pack up, my bag, -⟩ ⟨I, drive, car, -⟩

MIXED: An event that is rarely neutral, but is often considered positive by some people and negative by others. Both positive and negative understandings are common, and it is not clear which one is more dominant. For example, ⟨I, get, ticket, -⟩ can be positive in some contexts (e.g., getting tickets to a concert or sports games) and negative in some other contexts (e.g., getting tickets for speeding or an expired parking meter). Both of these situations are common, but the event is rarely neutral.

The first part of the research in this dissertation focuses on recognizing prior polarity of events, which is a stereotypical understanding and independent of context. Therefore, during the annotation stage, human annotators are required to assign a stereotypical polarity label to each event based on a general understanding of the event, rather than any special situation. For example, given the event ⟨we, go, -, to Disney World⟩, the stereotypical understanding is that people typically enjoy going to Disney World and usually have a lot of fun there. Thus, this event typically affects people positively and should be labeled as positive, even though there will always be exceptions (e.g., someone may have a bad experience there because the weather is not good or encounters somebody that the person dislikes). However, annotators were instructed to label events based on the norm, rather than exceptions.

3.3.2 Manual Annotation Study

To obtain gold standard annotations with affective polarity labels, three annotation volunteers (including the author) were instructed to assign one of the Positive, Negative, Neutral, and Mixed labels to each given event based on the definitions presented in the previous section. First, I randomly selected 1,500 events from the AffectEvent data set, and each annotator was required to annotate these events independently. After all annotators finished their annotations, I measured the pairwise inter-annotator agreement (IAA) scores

among the three annotators using Cohen’s kappa (κ) on the 1,500 events. Annotators achieved good annotator agreement scores on this task, which are $\kappa=.76$, $\kappa=.70$, and $\kappa=.69$.

To obtain a single annotator label for each event, I assigned the majority label to each event as the gold standard polarity. Only one event was labeled as Mixed, so I concluded that mixed polarity events are rare. Therefore, I abandoned the Mixed category and discarded the 1 Mixed event for the remainder of this research. In addition, there are 9 events that received three different labels from annotators, which did not have majority labels and were discarded too. Finally, the annotation process resulted in a gold standard data set of 1,490 events labeled as POSITIVE, NEGATIVE, or NEUTRAL. Of these 1,490 manually annotated events, I randomly selected 1,000 as the *test set* for evaluation and use the remaining 490 events as a *development set* for tuning parameters during my dissertation research. Some examples of the annotated events are shown in Table 3.1. More examples can be found in the Appendix B.

The distribution of polarities is shown in Table 3.2, which demonstrates that 38% of the randomly selected events have a positive or negative affective polarity, with slightly more positive events. These results suggest that affective events are pervasive, comprising nearly 4 of every 10 events, illustrating the importance of being able to recognize affective polarity of events for narrative text understanding.

Table 3.1: Examples of Gold Standard Affective Events

POSITIVE:	⟨I, play, music, -⟩
⟨kid, look up, -, to me⟩	⟨cost, be, low, -⟩
⟨I, go, -, to block party⟩	⟨someone, save, me, -⟩
⟨my confidence, rise, -, -⟩	⟨I, attend, show, -⟩
⟨I, dance, -, with my friend⟩	⟨I, kiss, her, -⟩
NEGATIVE:	⟨girl, laugh, -, at me⟩
⟨I, get, -, into argument⟩	⟨I, be, bummed, -⟩
⟨I, drop, my phone, in toilet⟩	⟨dog, pass away, -, -⟩
⟨house phone, not work, -, -⟩	⟨my face, look, pale, -⟩
⟨I, wake up, -, at 3 am⟩	⟨tear, pour, -, from eye⟩
NEUTRAL:	⟨I, pack up, my bag, -⟩
⟨I, decide to rent, car, -⟩	⟨trunk, be, open, -⟩
⟨tour bus, pull up, -, -⟩	⟨I, scribble, -, -⟩
⟨I, read, -, over post⟩	⟨I, have, staple, -⟩
⟨I, wake up, -, around 6⟩	⟨I, look, -, at sentence⟩

Table 3.2: The Distribution of Affective Events in the Gold Standard Data

POSITIVE	NEGATIVE	NEUTRAL
295 (20%)	264 (18%)	931 (62%)

Figure 3.2 shows the confusions among three annotators (i.e., Ann1, Ann2, and Ann3). The category labels are abbreviated as: positive (POS), negative (NEG), neutral (NEU), and mixed (MIX). #Tot denotes the total number of events in each row or column. The confusions among three annotators show that human annotators have difficulty in distinguishing positive or negative events from neutral, but they are good at recognizing positive events from negative ones.

3.4 Chapter Summary

This chapter presents a personal blogs corpus, and describes event frame representations and methods used for extracting event frames. This chapter also presents a manual annotation study to obtain gold standard data for evaluation and determine the prevalence of affective events. The content of this chapter is summarized below.

- This research created a *personal blogs corpus* that consists of 1,383,425 personal story blogs that are extracted from the data sets of the ICWSM 2009 and 2011 data chal-

Ann1. \ Ann2	POS	NEG	NEU	MIX	#Tot	Ann1. \ Ann3	POS	NEG	NEU	MIX	#Tot
POS	215	4	97	0	316	POS	261	10	45	0	316
NEG	11	203	62	0	276	NEG	3	247	26	0	276
NEU	32	23	840	1	896	NEU	47	56	793	0	896
MIX	2	2	7	1	12	MIX	2	7	2	1	12
#Tot	260	232	1006	2	1500	#Tot	313	320	866	1	1500

(a) Annotator 1 and 2

Ann2. \ Ann3	POS	NEG	NEU	MIX	#Tot
POS	219	14	27	0	260
NEG	2	210	20	0	232
NEU	92	96	818	0	1006
MIX	0	0	1	1	2
#Tot	313	320	866	1	1500

(b) Annotator 1 and 3

(c) Annotator 2 and 3

Figure 3.2: Polarity Annotation Confusions between Each Pair of Annotators.

lenges¹.

- This research designed two event frame representations: *basic event frame* and *enhanced event frame*. The basic event frame can only capture the agent, predicate, and theme of an event, and each component is represented with a single word. Compared with the basic event frame, the enhanced event frame contains more information. First, the enhanced event frame contains an additional prepositional phrase component. Second, the Agent, Theme, and Prepositional Phrase components in an enhanced event frame are composed with a minimal noun phrase (e.g., Named Entity or noun compound) rather than a single word. Third, the Theme component can be an adjective. Fourth, the predicate can be associated with a particle and an infinitive verb or gerund. Finally, active and passive event constructions are normalized using the enhanced event frame. The research presented in Section 4 uses the basic event frame representation. The research in Section 5 and 6 uses the enhanced event frame. Both basic and enhanced event frames are extracted using dependency relations.
- This chapter also presents a manual annotation study to obtain gold standard data of events with affective polarity labels, and determine the prevalence of affective events. With this annotation study, I obtained 1,490 events with Positive, Negative, and Neutral polarity labels, which are used as gold standard data for evaluation and are freely available to the research community. In addition, the manual analysis shows that 38% of events are affective (positive or negative), which suggests that affective events are pervasive, and illustrates the importance of recognizing affective events for narrative text understanding.

¹The ICWSM 2009 data can be found here: www.icwsm.org/2009/data/, and the ICWSM 2011 data can be found using this link: www.icwsm.org/data/.

CHAPTER 4

EXTRACTING AFFECTIVE EVENTS FROM BLOGS WITH EVENT CONTEXT GRAPH MODEL

In this chapter, I present the details of a semi-supervised Event Context Graph (ECG) model for identifying events that are typically associated with a positive or negative state from a large collection of personal blogs. First, I extract event representations from personal story blogs using the basic event frame representation described in Chapter 3, and then construct an enormous event context graph that contains nodes representing events and sentences in the corpus. This model is based on two intuitions: (1) some instances of affective events will occur in contexts mentioning explicit sentiment indicators (e.g., “good”, “bad”, “happy”, or “upset”), and (2) events that frequently co-occur in the same documents tend to share similar affective polarities. The ECG model links event mentions with their contexts and connects event pairs that frequently co-occur to allow affective evidence to propagate from contexts to events and among events. The model explores three types of edges: *event-sentence* edges capture local context, *sentence-sentence* edges capture discourse proximity context, and *event-event* edges capture event co-occurrence in documents. With the event context graph, the model first assigns initial affective polarities with a sentiment classifier to “seed sentences” and uses a semi-supervised label propagation algorithm to extract affective events by spreading affective evidence across edges of the graph.

In the rest of this chapter, I first present the motivation, present the overview of the Event Context Graph model, and describe the details of each part of the model including a sentiment sentence classifier used for identifying seeding sentences, an event context graph, and a label propagation algorithm. Then, I describe baseline methods using existing sentiment resources and evaluations on manually labeled data sets. Finally, I summarize

this chapter.

4.1 Motivation

Supervised learning can be used to recognize affective events, but it usually requires a large amount of human-labeled data to achieve good performance, which is time consuming and expensive. This chapter presents a semi-supervised Event Context Graph (ECG) model, which has the advantage that it only needs automatically obtained noisy labels and can be adapted to other data sets with relatively less effort. The ECG model is motivated by the following observations and intuitions. First, events are extracted from personal story blogs in which 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. I observed that when bloggers describe events that happened to them, they sometimes explicitly express their sentiments toward the events in contextual sentences. In some cases, the explicit sentiment is expressed in the same sentence where an event is mentioned (e.g., “I’m so happy that I’m graduating”). For some other cases, the explicit sentiment is described in nearby sentences. For example, in the description “It was so fun yesterday. I went to Disney World with my family”, the blogger expressed a positive sentiment in the first sentence, and described the corresponding event in the second sentence. Based on this observation, I explore 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 whose polarities can potentially be recognized using existing sentiment analysis tools. In addition, I noticed that events that frequently co-occur in the same documents tend to have the same affective polarity. For example, we can expect that the events “John was hit by a car” and “John was sent to the hospital” will frequently co-occur in the same blogs. Based on this observation, if we can determine the polarity of an event based on its discourse contexts, then we can also spread this polarity evidence to other highly correlated events. For example, if we know “John was hit by a car” is negative, we can reasonably infer that “John was sent to the hospital” is also negative. However, a key challenge of this research is to explore the effectiveness of different types of discourse contexts and collocation metrics in learning to recognize events that have strong affective polarities. In the following sections, I will

present the details of the Event Context Graph model.

4.2 Semi-Supervised Learning of Affective Polarity with Event Context Graph Model

4.2.1 Overview

To learn affective events, I begin by extracting frequent events from a large set of personal story blog posts. Then, I identify affective events from the frequent events using semi-supervised learning with an Event Context Graph (ECG) model. An event context graph is first created, which consists of two types of nodes: event nodes and sentence nodes, and three types of edges: *event-sentence* edges capture local context, *sentence-sentence* edges capture discourse proximity context, and *event-event* edges capture event co-occurrence in blog posts. To initialize the ECG model, I apply a sentiment sentence classifier to identify sentences that have strong positive or negative polarity, which become the seed nodes for semi-supervised learning. Then, I use a label propagation algorithm to iteratively spread polarity evidence across the edges of the event context graph from the seed nodes. Consequently, events that frequently occur in affective contexts will be assigned high values for affective polarity through *event-sentence* edges. With *sentence-sentence* edges, the propagation algorithm can spread affective evidence across local discourse regions, and *event-event* co-occurrence edges will propagate affective evidence from some events to other frequently collocated events. Finally, each event will be assigned a affective polarity value, which can be used to extract strongly affective events.

4.2.2 Sentiment Sentence Classifier

Existing sentiment analysis resources are not adequate to recognize many affective events, but they can identify affective polarities of many sentences that contain explicit sentiment indicators. My research exploits these resources to automatically obtain noisy supervision for the semi-supervised Event Context Graph model. Specifically, I created and applied a sentiment sentence classifier to identify an initial set of sentences that have strong positive or negative polarities. The identified sentences were used as seeding nodes for the label propagation algorithm. This section provides details of building a sentiment sentence classifier.

I created a logistic regression classifier that labels sentences as having *positive*, *negative*,

or *neutral* polarity. This classifier provides a probability value that is used to assign a confidence strength to each label. The features for this classifier were designed to be similar to that used in the NRC-Canada sentiment classifier (Mohammad et al., 2013) which performed very well in the SemEval 2013 Task 2 (i.e., the Sentiment Analysis Task in Twitter). Since the blog data used in this research is also a form of social media text, this feature set was expected to be also well-suited for the blog data. I did not include all of features designed for the NRC tweet sentiment classifier because some features are specific to tweets rather than blogs (e.g., elongated words and emoticons). The sentiment classifier was trained on tweets with gold affective polarity labels, and features were extracted for each tweet during the training stage. When the classifier was applied to blog posts to determine the polarity for an input sentence, features were extracted from the sentence. Specifically, the sentiment classifier was built using following features.

- **Word N-grams:** For a tokenized text, I first convert words in the text into lower case and then create features for each unigram, bigram, trigram, and 4-gram. Each n-gram feature is assigned with a binary presence value, i.e., the feature value is set to 1 if the n-gram appears in the given text, otherwise 0.
- **Word Skip-grams:** To obtain these features, I first obtain the trigrams and 4-grams from a tokenized text (all letters are in lowercase). Then, I extract word ski-gram features by replacing one word of a trigram or 4-gram with an asterisk *. The values of these features are binary.
- **Character N-grams:** For a given text string (in lowercase), I create features for each contiguous sequence of 3, 4, and 5 characters. Each feature is assigned with a binary presence value.
- **Capitalization:** For an input text, this feature counts the total number (n) of words in which all letters are in uppercase. The feature value is set to n .
- **Part-Of-Speech:** To obtain this type of features, I first obtain part-of-speech tags for an input text. Then, I create a feature for each unique part-of-speech tag. The value for a part-of-speech tag feature is set to the total number of times the tag appears in the text.

- **Hashtag:** I also create a feature to indicate how many hashtags are in a text. The feature value is set to the total number of any hashtags mentioned in the text.
- **Sentiment Lexicons:** this type of features were extracted from six sentiment lexicons: MPQA Subjectivity lexicon (Wilson et al., 2005), Hu & Liu’s lexicon (Hu and Liu, 2004), NRC Emotion lexicon (Mohammad and Turney, 2010), NRC Hashtag lexicon (Mohammad et al., 2013), Sentiment140 lexicon (Go et al., 2009), and AFINN lexicon (Nielsen, 2011). Some lexicons consist of individual words, while others contain bigrams and word pairs. Therefore, to match a lexicon, I first convert an input text into a list of terms, which could be individual words, bigrams, or word pairs depending on the lexicon. Then, for each lexicon, I create the following 7 features using the prior polarity scores (which could be binary or real value scores) of terms in the lexicon. (1) I create a feature to count the number of terms whose polarity scores are greater than 0 based on the lexicon. (2) For a given text, I also create a feature to denote the sum of the polarity scores for all terms in the text, (3) the sum of positive scores (i.e., score > 0) for all terms in the text, (4) the sum of negative scores, (5) the maximal score for all terms in a given text, (6) the minimal score for in the text, (7) the polarity score of the last term in the text.

With the above features, I built a logistic regression classifier using the tweet data with three polarity labels (positive, negative, and neutral) from the SemEval 2014 Task 9 (Sentiment Analysis Task in Twitter), in which 6425 tweets were used for training and 1564 tweets were used for evaluating the performance of the classifier. The classifier’s macro average performance is 69.4% Precision, 68.2% Recall, and 67.8% F1, which is close to the performance (which is 69.02% F1) reported for the NRC-Canada system for the SemEval 2013 Message-level Task (Mohammad et al., 2013).

4.2.3 Event Context Graphs

To identify affective events, I construct an *Event Context Graph* that links events with their contexts and other co-occurring events. The graph contains two types of nodes: event nodes and sentence nodes. Each event node denotes a unique event frame extracted from the blog corpus, and each sentence node represents a sentence from a blog post. Nodes are connected with three types of edges: *event-sentence*, *sentence-sentence*, and *event-event*

edges, which are created to investigate different ways of propagating polarity evidence between events and contexts. Figure 4.1 shows an illustration of an Event Context Graph. The following describes how these edges are created.

- Event-Sentence Edges:** These edges connect events with sentences in which they appear. With event-sentence edges, the affective polarity of an event can be induced from its local contexts. For example, if the event $\langle I, \text{graduate}, - \rangle$ is extracted from the sentence “I’m so happy that I’m graduating”, the event and sentence will be linked together. The polarity evidence in the sentence, which originates from the word “happy”, will be propagated to the event through this connection. Since an event may appear in multiple contexts, its affective polarity is induced using all of the contexts in which it occurs. When a sentence node s_i is linked to an event node e_j , the weight on this edge 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 , i.e., all the unique events that are mentioned in the sentence s_i . Intuitively, this assumes that events mentioned in sentence s_i contribute equally to the polarity of s_i .
- Sentence-Sentence Edges:** These edges connect adjacent sentences in the same documents. They allow for label propagation across neighboring sentences to capture the intuition that sentences in the same discourse region are likely to have the same polarity. For example, given two adjacent sentences “(1) Last weekend was so fun. (2) We had a campfire at the beach.”, the edge between these two sentences will allow for the propagation of the explicit polarity evidence from the first sentence to the second sentence. I set the weight for an edge linking sentence s_i and s_j to be $w(s_i, s_j) = 0.80$, which indicates that I expect the sentiment evidence expressed in

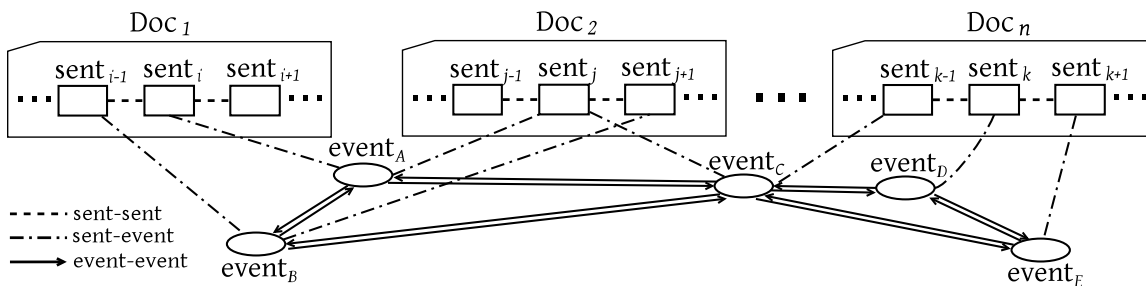


Figure 4.1: Illustration of an Event Context Graph with Three Types of Edges

nearby contexts will gradually degrade based on the distance to the current sentence when it is aggregated into the current sentence. With the weight of 0.8, the sentiment evidence value (suppose the original value is 1.0) of a sentence s will degrade to 0.8 (i.e., $1.0 * 0.8 = 0.8$) when it is passed to sentences with distance 1 to sentence s (i.e., the previous or following sentence), and 0.64 (i.e., $1.0 * 0.8 * 0.8 = 0.64$) when it reaches sentences with distance 2 to sentence s . However, it will degrade to around 0.3 when it is passed to sentences of distance 5 (i.e., $1.0 * 0.8^5$), which indicates that sentiment evidence from contextual sentences will have little influence when the context window is ≥ 5 . In my experiments, I intuitively chose this value based on the above analysis, and did not experiment with other values, so exploring methods to find an optimal weighting could be fruitful in future work.

- **Event-Event Edges:** These edges link events that frequently co-occur in the same blog post, and are designed to capture the intuition that if two events frequently co-occur in same stories, then they are likely to share the same affective polarity. For example, we would expect the events $\langle \text{kid, be_hurt, -} \rangle$ and $\langle \text{kid, cry, -} \rangle$ to frequently co-occur in blog posts by parents discussing accidents involving their children, and they have the same negative polarity. For this type of edge, I 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:

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

where $p(e_i, e_j)$ is the probability that events e_i and e_j appear in the same document, and $p(e_i)$ is the probability that e_i appear in the personal blog corpus.

4.2.4 Variants of the Event Context Graph Model

To investigate the effectiveness of each type of edge, I create three graph variants by incrementally incorporating *event-sentence* edges, *sentence-sentence* edges, and *event-event* edges. Each of the three graph architectures is described below.

- **Local Context Graph (G^{LOC}):** This graph configuration contains only *event-sentence* edges. Affective polarity of an event can only be obtained from sentences that mention the event.

- **Discourse Context Graph (G^{DIS}):** This graph configuration contains two types of edges: *event-sentence* edges as well as *sentence-sentence* edges. In this variant, affective polarity evidence can propagate across sentence boundaries as well. Events can receive polarity evidence not only from its local context but also from the discourse context.
- **Event Co-Occurrence Graph (G^{EV}):** This graph configuration contains three types of edges: *event-sentence*, *sentence-sentence*, and *event-event* edges. Besides local and discourse context, affective polarity evidence can also be propagated through connections between co-occurring events. This is the full Event Context Graph (ECG) model.

4.2.5 Semi-Supervised Label Propagation

To learn affective polarities of events, I use the label propagation algorithm from (Zhu and Ghahramani, 2002). The pseudocode for my implementation is shown in Algorithm 1, which will outputs a real valued polarity score ($\psi(v) \in [-1, +1]$) associated with each node. The polarity value of -1 denotes a negative polarity, +1 means a positive polarity, and 0 denotes the node is neutral.

Algorithm 1 Semi-Supervised Label Propagation

Require: $G(V, E)$, Sentiment Classifier SC , seed threshold τ

Ensure: $\psi(v) \in [-1, +1], \forall v \in V$

- 1: Initialize seed nodes using SC with threshold τ
 - 2: **while** ψ has not converged **do**
 - 3: Update $\psi(v)$ using Equation 4.1.
 - 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 starts with a small set of manually selected seed words. Instead, I initialize the label propagation algorithm with a set of sentence contexts that are determined to be positive or negative. I apply a sentence classifier designed

for sentiment analysis (described in Section 4.2.2) to all of the sentences in the blogs and identify sentences that are classified as having positive or negative polarity. The highly confident positive or negative sentences are then used as seeding information for the Event Context Graph model.

The sentiment classifier is not perfect (69% precision), and the initial set of labeled nodes need to be as accurate as possible. Therefore, I only select a sentence node as seeding 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 $\geq \tau$ by the sentiment classifier in the initialization step of Algorithm 1. In the experiments, τ is set to be 0.5. Since the sentiment classifier was trained on data with three polarity classes, the classification decision with probability over 0.5 was fairly accurate (which was 77% precision based on tweet evaluation data). 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.

$$\psi^{t+1}(v) = \frac{\sum_{v' \in N(v)} w(v, v') * \psi^t(v')}{\sum_{v' \in N(v)} w(v, v')} \quad (4.1)$$

After initialization, the affective polarity values are iteratively propagated across edges. I compute the affective polarity $\psi(v)$ of node v as the weighted average of the polarity values of its neighbor nodes $N(v)$. Formally, I use Equation 4.1 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. The label propagation algorithm is guaranteed to converged (Zhu and Ghahramani, 2002). In my experiments, I ran label propagation until it ran for 100 iterations.

4.3 Evaluation

This section presents evaluation details of the Event Context Graph model described in the above section. For comparison, I designed two baseline systems that took advantage of local contexts (sentences) and global contexts (documents). This section first describes the two baseline systems, and then presents the evaluation data, metrics, and experimental

results.

4.3.1 Baseline Systems

For comparison, I designed two baseline systems that acquire affective events by applying a sentiment classifier to the context surrounding events. The first system, **AvgSent**, is designed to identify affective events that frequently occur with an explicitly expressed sentiment. For example, the event $\langle \text{we, have, campfire} \rangle$ is typically a fun experience, so I expect to find sentences such as “We will have a campfire, so excited!”. For each event e , the **AvgSent** system collects all of the sentences containing the event e and applies the sentiment classifier described in Section 4.2.2 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 , and $p(s)$ 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 overall positive or negative polarities. For example, bloggers sometimes express sentiment at the beginning of their post (e.g., “So happy today!”), after which they describe the events that happened to them. 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 the same polarity, then the event is likely to have that polarity. For each event 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:

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

where $D(e)$ is the set of documents containing event e , and $p(s)$ is still the signed polarity score of sentence s .

4.3.2 Evaluation Data Set and Metrics

For the evaluation presented in this section, I used the blog posts from the ICWSM 2009 Spinn3r data set. After applying the preprocessing steps described in Section 3.1, I obtained around 1.4 million personal story blog posts, which were automatically processed using Stanford CoreNLP tools and parsed using the Stanford dependency parser. I used the basic event frame representation described in Section 3.2 to collect event representations from the personal blogs. To keep the size of the graphs manageable, events with frequency less than 50 in this data set were filtered. This produced 40,608 unique basic event frames.

In these experiments, I evaluated the affective events produced by the AvgSent and AvgDoc baseline systems as well as the three variants (i.e., G^{LOC} , G^{DIS} , G^{EV}) of the Event Context Graph with label propagation. After collecting event representations from the personal blogs, I applied each of the five methods and ranked the events based on the affective polarity scores produced by that method. Because both positive and negative values can denote strong affective events, I used the absolute values of the polarity scores to generate a single ranking of events with both positive and negative polarity.

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, I applied the sentiment classifier to each event and separated out the events that the classifier labeled as positive or negative. I 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 (because the sentences containing these events are likely to have been seed nodes). Second, the goal of this research was to learn *implicitly* affective events. Since manual annotation is expensive, I wanted to focus the 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, I evaluated the precision of the top-ranked k affective events hypothesized by each method. For each of the five systems, I 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.

I 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. The definition of each label is shown below.

- **Positive:** The event is typically desirable or beneficial. For example: ⟨I, see, sunset⟩
- **Negative:** The event is typically undesirable or detrimental. For example: ⟨girl, have, flu⟩
- **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: ⟨he, open, door⟩
- **Invalid:** The event does not describe a sensible event. This label is primarily for erroneous event representations resulting from preprocessing or parsing mistakes. For example: ⟨cant, do,⟩

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

4.3.3 Experimental Results

Table 4.1 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 not perform as well, suggesting

Table 4.1: Accuracy for the Top-Ranked Affective Events.

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

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 accuracy 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 accuracy 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% accuracy for the top 100 events and slightly higher accuracy (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 many affective events that a sentiment classifier did not recognize as having polarity.

4.4 Analysis

In this section, I analyze the performance of the ECG model. First, I show some examples of the top-ranked events produced by the G^{EV} system, and analyze events that were incorrectly recognized as affective. Second, I further investigate the ability of existing sentiment lexicons to identify affective events from the top-ranked events produced by the G^{EV} system. Finally, I evaluate the ECG model on another set of randomly sampled events for comparison with the model to be described in Chapter 5.

4.4.1 Analysis of Learned Affective Events

Table 4.2 shows the top 50 positive events and the top 50 negative events produced by label propagation with the G^{EV} graph.

Table 4.2: Top 50 Positive and 50 Negative Affective Events Produced with Label Propagation with G^{EV} , \emptyset Denotes Empty Element. Verbs that Usually Occur with a Particle Are Denoted with * (e.g., *screw up*, *black out*, *shut down*) to Help Readers Interpret the Likely Intended Phrase.

Negative Events		
⟨he, lose, mind⟩	⟨she, lose, lot⟩	⟨I, break, nose⟩
⟨phone, be_broken, \emptyset ⟩	⟨professional, advise, \emptyset ⟩	⟨life, lose, \emptyset ⟩
⟨she, lose, balance⟩	⟨he, lose, balance⟩	⟨phone, break, \emptyset ⟩
⟨Im, stick, \emptyset ⟩	⟨tear, sting, eye⟩	⟨she, be_hit, by_car⟩
⟨I, fall, bike⟩	⟨he, lose, lot⟩	⟨she, hit, head⟩
⟨Im, screw, \emptyset ⟩*	⟨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, \emptyset ⟩
⟨he, have, seizure⟩	⟨neck, start, \emptyset ⟩	⟨I, sprain, ankle⟩
⟨I, cut, finger⟩	⟨he, hit, head⟩	⟨heart, start, pound⟩
⟨he, lose, her⟩	⟨she, stick, head⟩	⟨she, lose, track⟩
⟨head, pound, \emptyset ⟩	⟨she, lose, control⟩	⟨bone, be_broken,⟩
⟨he, lose, job⟩	⟨I, seethe, \emptyset ⟩	⟨I, break, bone⟩
⟨she, stick, hand⟩	⟨I, screw, thing⟩*	⟨I, injure, myself⟩
⟨time, lose, \emptyset ⟩	⟨she, black, \emptyset ⟩*	⟨I, be_stung, by_bee⟩
⟨he, be_rushed, \emptyset ⟩	⟨I, call, vet⟩	
Positive Events		
⟨we, sing, birthday⟩	⟨all, have, weekend⟩	⟨everyone, have, weekend⟩
⟨learn, make, money⟩	⟨learn, use, \emptyset ⟩	⟨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, \emptyset ⟩
⟨time, be_had, \emptyset ⟩	⟨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, \emptyset ⟩	⟨we, have, shopping⟩
⟨we, have, dancing⟩	⟨we, see, view⟩	⟨i, see, sunset⟩
⟨we, have, time⟩	⟨kid, have, lot⟩	⟨we, reunite, \emptyset ⟩
⟨you, go, girl⟩	⟨i, find, deal⟩	⟨we, laugh, lot⟩
⟨we, have, turn⟩	⟨all, have, time⟩	⟨god, have, sense⟩
⟨we, have, weather⟩	⟨it, entertain, \emptyset ⟩	⟨we, have, feast⟩
⟨i, get, present⟩	⟨we, have, cookout⟩	

Negative events include many physical injuries and ailments, car accidents, and lost or broken phones. Note that $\langle I, \text{call}, \text{vet} \rangle$, $\langle \text{she}, \text{black}, \emptyset \rangle$, and $\langle \text{he}, \text{be_rushed}, \emptyset \rangle$ often suggest medical emergencies (i.e., “she blacked out” and “he was rushed to the hospital”). Since these events were extracted using the basic event frame representation, some event components were not precisely extracted, which can be solved using the enhanced event frame representation that is used in Chapters 5 and 6. Positive events include birthdays, playing, dancing, shopping, and cookouts. There are also more subtle examples of stereotypically enjoyable situations, such as $\langle \text{we}, \text{find}, \text{deal} \rangle$, $\langle \text{we}, \text{make}, \text{team} \rangle$, $\langle I, \text{see}, \text{sunset} \rangle$, $\langle \text{we}, \text{reunite}, \emptyset \rangle$, and $\langle \text{room}, \text{have}, \text{view} \rangle$.

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, $\langle \text{we}, \text{have}, \text{weather} \rangle$ is not a positive or negative event per se, but originates from sentiments expressed about the weather, such as “we have nice weather”. These results suggest that people use this expression to comment a good weather more than a bad weather. Similarly, $\langle \text{it}, \text{leave}, \text{taste} \rangle$ isn’t negative per se, but comes from the common expression “it leaves a bad taste”.

4.4.2 Performance of Sentiment Lexicons

As a reminder, none of the events shown in Table 4.2 were identified as having positive or negative polarity by the sentiment classifier described in Section 4.2.2. However, I further explored whether sentiment lexicons can recognize the affective polarity of these events. I used four well-known sentiment/opinion lexicons: ConnotationWordNet (Kang et al., 2014), MPQA Subjectivity Lexicon (Wilson et al., 2005), SenticNet3.0 (Cambria et al., 2014), and SentiWordNet3.0 (Baccianella et al., 2010). For each event, I assigned an affective polarity based on the polarities of its component words. For an event e , I computed a polarity score using the following formula:

$$s(e) = \frac{1}{n} \sum_{i=1}^n \text{lex_score}(w_i)$$

where n is the total number of words in the event representation, w_i is a word with the index i , and $\text{lex_score}(w_i)$ is the polarity score given by the lexicon. If the word w_i is not in the lexicon, I use a default value of 0, which denotes that the word has no bias toward

positive or negative polarity. I also look for the presence of negation, and multiply the score by -1 if negation is found.

Since MPQA provides discrete labels, I 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 I 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 I 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 I assigned polarity scores as: *negative* if $s(e) \in [0, .5 - \lambda)$, *positive* if $s(e) \in (.5 + \lambda, +1]$, and *neutral* otherwise. I experimented with λ values ranging from 0 to 0.4 in increments of 0.1 to identify the most effective setting for each lexicon. Table 4.3 shows the results for $\lambda = 0$ and the lambda value producing the best precision on the positive class for each lexicon. I also tried to select lambda values based on the precision for both positive and negative, but the results are similar.

Table 4.3 shows the results for each sentiment lexicon as well as label propagation with our Event Context Graph G^{EV} , in which #Events denotes the number of correctly recognized events. I 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, it can be seen 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

Table 4.3: Evaluation of Polarity Labels Assigned by Label Propagation with G^{EV} and Four Sentiment Lexicons

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

events, ranging from 31% for the most prolific output (CWN with $\lambda=0$) 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 it failed to recognize many of the 377 negative events discovered by G^{EV} .

Overall, this analysis reveals that many affective events are not recognized as having polarity by these sentiment lexicons. 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 approach.

4.4.3 Evaluation on Randomly Sampled Events

In the experiments described in previous sections, events were extracted using the basic event frame representation from only the ICWSM 2009 data, and the experiments only evaluated the top-ranked affective events that were not identified as having positive or negative polarity by the sentiment classifier. This section further investigates the performance of the Event Context Graph model on a set of randomly selected events.

In this evaluation, I used the personal blogs corpus from both the ICWSM 2009 and 2011 data sets (described in Section 3.1), which consists of nearly 1.4 million story blog posts, and the gold annotations described in Section 3.3. To capture richer event phrases, I used the enhanced event representation and extracted 571,424 unique events that have frequency of at least 5 in the corpus. Then, I applied the Event Context Graph (ECG) model (G^{EV}) to learn affective events. The ECG model produces real polarity scores ranging from $[-1, +1]$, rather than discrete category labels. Therefore, I used the same method as that used for the sentiment lexicons to convert real polarity scores to discrete polarity classes. Specifically, I assigned the polarity label of an event as: *negative* if its polarity score is in $[-1, -\lambda)$, *positive* if the score is in $(+\lambda, +1]$, and *neutral* otherwise. The parameter λ can take any value between +1 and -1. The gold data contains two sets: the test set of 1000 events, and the development set of 490 events. I tuned a λ parameter on the development data by choosing values ranging from 0.05 to 0.5 in increments of 0.05, and found the best value to be $\lambda = 0.15$.

Table 4.4 shows the precision (Pre), recall (Rec), and F1 scores for each polarity class, and the overall average F1 score (AvgF1). The top portion of Table 4.4 shows the results on the development set with different λ values. For both positive and negative polarities, the precision increases with increasing λ values, but the recall decreases. For example, with $\lambda = 0.45$, the model achieved 100% and 83.3% precision on positive and negative events, but only 5.2% and 5.7% recall, respectively. The precision on positive events is higher than negative, but the recall is lower. Overall, the ECG model obtained the best AvgF1 score of 46.6% with $\lambda = 0.15$. The bottom portion shows the performance on the test set with the best $\lambda = 0.15$, which was selected on the development set. It can be seen that the model produced more negative events (67.8% recall) than positive events, but the precision is not satisfying. As one would expect, the precision of affective events on randomly sampled events is lower than that for the top-ranked events reported in the previous section. This demonstrates that more effective methods are needed to improve the accuracy of identifying affective events more generally.

4.4.4 Discussion of Limitations

This section discusses the limitations of the Event Context (ECG) graph model from the perspective of performance and efficiency.

Besides evaluating the Event Context Graph model on the top-ranked events, I also looked at the polarity values of other events, which were produced by the model. I noticed that the polarity values of many affective events are very close to 0, which means that the affective event is not recognized. This observation can also be verified by the results

Table 4.4: Performance of ECG Model on Randomly Sampled Events

λ	POSITIVE			NEGATIVE			NEUTRAL			AvgF1
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	
Development Set										
0.05	43.1	28.9	34.6	22.7	88.5	36.2	72.1	20.3	31.6	34.1
0.15	70.0	14.1	23.9	34.5	66.7	45.5	70.5	69.6	70.1	46.5
0.25	88.9	8.2	15.1	54.8	26.4	35.7	66.1	94.8	77.9	42.9
0.35	100	7.2	13.5	66.7	13.8	22.9	64.7	98.4	78.1	38.1
0.45	100	5.2	9.8	83.3	5.7	10.8	63.7	99.7	77.7	32.8
Test Set										
0.15	68.6	17.7	28.1	34.9	67.8	46.1	66.9	64.8	65.9	46.7

presented in Table 4.4 on a randomly sampled events, in which the recall for positive polarity is 8.2% and 26.4% for negative polarity when setting the threshold $\lambda = 0.25$. This reveals that, for many affective events, the affective polarity evidence is not propagated to them during the learning process. This can be caused by two potential problems. First, the seeding information is not adequate. In the ECG model, I first obtained polarity predictions for sentences in the data using a sentiment classifier and then selected sentences with highly confident polarity predictions as seeding nodes, which is the only supervision for the ECG model. For the label propagation algorithm, if a node is not a seeding node, then it is initialized with default value 0. Therefore, in the ECG model, if an event's discourse contexts have no seeding nodes, and the contextual sentences of its co-occurring events have no seeding nodes either, then the polarity evidence will not be propagated to this event. Second, the edges between events and its contextual sentences, and other co-occurring events are not adequate. Without sufficient connections between an event and other resources, it is difficult for polarity evidence to be propagated to this event.

For efficiency, I will first analyze the time complexity of the label propagation algorithm. Let $G = (V, E)$ denote a graph, in which V is the set of graph nodes and E denotes the set of graph edges. To estimate the time complexity, I first compute the total number of basic operations performed at each iteration, which can be measured by counting arithmetic operations in Equation 4.1. For a given node v_i and its neighbors $N(v_i)$, the number of operations in the numerator is the sum of $|N(v_i)|$ multiplications and $|N(v_i)| - 1$ additions. There are also $|N(v_i)| - 1$ additions in the denominator. Adding another 1 division, there are $3|N(v_i)| - 1$ operations for event v_i at each iteration. By summing up the number of operations for each event, there are a total of $\sum_{v_i \in V} 3|N(v_i)| - 1$ operations at one iteration. Since the number of neighbors of an event is equal to the number of edges connected to it, summing up the number of neighbors for all events is equal to counting each edge twice when the edges are undirected. In the ECG model, some edges are directed. Therefore, the total number of operations can be approximated as $\sum_{v_i \in V} 3|N(v_i)| \leq 2|E|$ where $|E|$ is the number of edges in the graph. Therefore, the time complexity of the label propagation is $O(|E|)$. Based on this analysis, we can see that the model's efficiency is closely related to the number of edges in the graph.

To investigate how event nodes and sentence context nodes contribute to the complex-

ity of the model, I analyzed the numbers of the two types of nodes in the Event Context Graphs. In the experiments described in Section 4.3, there are three graph variants. I obtained that, in total, the Local Context Graph (G^{LOC}) contained roughly 12 million nodes, and the Discourse Context Graph (G^{DIS}) and Event Co-Occurrence Graph (G^{EV}) each contained roughly 25 million nodes. However, there were only 40,608 event nodes in these graphs, which means that more than 99% of nodes are sentence nodes. The goal of this research is to extract affective events rather than sentences. However, the learning algorithm spends the majority of its computation on sentence nodes, which is inefficient. In addition, representing sentence contexts as nodes makes it difficult to apply the model on larger data sets because the large number of sentence nodes requires a lot of memory.

Overall, the above analysis demonstrates that more effective methods need to be explored to improve the performance of identifying affective events.

4.5 Chapter Summary

This chapter first describes the details of the Event Context Graph model, presents evaluation experiments, and then gives a further analysis of the model by comparing it with existing sentiment lexicons and evaluating on randomly sampled events. A discussion about the model’s efficiency and performance is presented at the end of this chapter. The content of this chapter is summarized below.

- This chapter presents an Event Context Graph model that was designed to extract affective events using discourse context and event collocation information. First, the model builds an Event Context Graph that consists of event nodes and sentence nodes, which are linked with *event-sentence*, *sentence-sentence*, and *event-event* co-occurrence edges. Then, a sentiment classifier is created to identify sentences with strong polarity values as seed nodes, and a label propagation algorithm is applied to iteratively spread polarity evidence from seed nodes to other unlabeled event and sentence nodes. Finally, affective events are extracted using the polarity scores learned by the label propagation algorithm.
- This chapter presents evaluations on a set of top-ranked events that were produced by three variants of the ECG model and two baseline systems, and manually anno-

tated by Amazon's Mechanical Turk workers. Experimental results show that the ECG model can recognize affective events missed by sentiment classifiers, and continually obtain better performance by increasingly adding local context, discourse context, and event co-occurrence information into the model.

- This chapter also investigates whether the learned affective events can be recognized by existing sentiment lexicons. I applied several well-known sentiment lexicons on the affective events produced by the best ECG model. The results demonstrate that sentiment lexicons failed to recognize many affective events discovered by the ECG model. In addition, this chapter evaluated the ECG model on randomly sampled events and discussed limitations of the model. The analysis shows that more efficient methods need to be explored to improve performance of identifying affective events.

CHAPTER 5

RECOGNIZING AFFECTIVE EVENTS USING SEMANTIC CONSISTENCY GRAPH MODEL

The previous chapter presents an Event Context Graph (ECG) model to extract affective events using event discourse and collocation information. Though the ECG model can identify many affective events, further analysis shows that more efficient and accurate methods are needed to identify larger sets of affective events. This chapter introduces a Semantic Consistency Graph (SCG) model that automatically induces a large collection of affective events from a personal story corpus with better efficiency and accuracy. In contrast to the ECG model, the SCG model tries to infer affective polarities of events by optimizing the semantic consistency in a graph rather than propagating polarity evidence from event contexts to events.

Briefly, the SCG model first extracts events from the personal story corpus using the enhanced event frame representation (described in Section 3.2), and then builds a graph consisting of nodes corresponding to events and their components. In the graph, event nodes are linked based on three types of semantic relations: (1) *semantic similarity*, (2) *semantic opposition*, and (3) *shared components*. Next, initial polarity values are assigned to events using existing sentiment analysis resources (e.g., sentiment lexicons and sentiment classifiers). Although sentiment resources are not very accurate for many affective events, they can recognize events that have explicitly affective language (e.g., “I had fun” or “I yelled in anger”). Consequently, the initialization step serves as noisy supervision. The learning algorithm is then designed to infer the true polarity values of events by iteratively refining the polarity values to optimize for the overall *semantic consistency* in the graph. Intuitively, the algorithm encourages semantically similar events to have similar polarity, semantically opposing events to have opposite polarity, and events to have polarity values consistent with their components. I applied this model to a corpus of nearly 1.4

million personal stories and induced a collection of $>110,000$ affective events with over 90% precision for positive events, and 80% for negative events. Experiments show that the SCG model achieved higher recall and precision on a set of randomly sampled events (described in Section 3.3) than previous affective lexicons and learning models.

In the following sections, I first present the motivation and design details of the SCG model, and then evaluate the model on a set of manually labeled events and compare it with previous methods. At the end, I will further analyze the behavior of the model by investigating the cases that were correctly and incorrectly classified, and also show the quality and quantity of affective events identified by the model.

5.1 Motivation

The Event Context Graph (ECG) model introduced in Chapter 4 was also designed to extract affective events from personal story blogs. One disadvantage of the ECG model is that it cannot be easily extended to a much larger data set because of the large number of sentence nodes in the Event Context Graph. The goal of the ECG model is to estimate the polarity of events rather than sentences, but as analyzed in Section 4.4, over 99% of nodes in the graph are sentence nodes, which adds a heavy computational burden to the model and makes it inflexible to extract more affective events from larger data sets. In addition, as shown in Section 4.4, the ECG model also failed to identify many affective events when evaluated on a set of randomly sampled events. Motivated by these two limitations of the previous ECG model, I explored new methods to extract more affective events with better efficiency and accuracy.

The Semantic Consistency Graph (SCG) model presented in this chapter is motivated by three intuitions that correspond to three types of semantic consistencies. The first intuition is that semantically similar events usually have similar affective polarities. For example, the events “I am unemployed” and “I lost my job” are semantically similar and they have the same negative polarity. The second intuition is that semantically opposite events often have opposite affective polarities. For instance, events “I won” and “I did not win” have opposite semantic meanings and their affective polarities are also opposite, i.e., the first is positive and the latter is negative. The third intuition is that the affective polarity of an event often originates from its components. For example, the event “I went

to a birthday party” is positive mainly because its component “birthday party” is positive. Based on these intuitions, the SCG model builds a graph to incorporate the three types of semantic relations and seeks to achieve a state where the affective polarities of events are consistent with their semantic relations. In addition, motivated by the observation that many affective events can be recognized by existing sentiment analysis resources even though some estimations are incorrect, I hypothesize that the polarity values obtained from existing sentiment analysis resources can be used as weak supervision to infer the true polarity of events.

5.2 Semantic Consistency Graph Model

5.2.1 Overview

The Semantic Consistency Graph (SCG) model is a weakly supervised method that was designed to automatically learn a large set of affective events. The key idea is to define a graph of events that are connected using different types of semantic relations, and to use the initial affective polarities estimated from sentiment analysis resources as noisy supervision. Using an optimization framework, the model can then learn the correct polarity values by enforcing semantic consistency across the semantic relations in the graph.

Figure 5.1 shows an illustration of the semantic relations graph used in the SCG model. The graph contains nodes corresponding to events (V_i) and components (C_k), which are the 4 composing parts (e.g., Agent, Predicate, Theme, and Prepositional Phrase) of an event representation. The graph also has three types of edges: *event-event similarity edges* that link semantically similar event pairs, *event-event opposition edges* (blue dotted line) that link semantically opposing event pairs, and *event-component edges* that connect an event with its components individually. The learning model will prefer that semantically similar events have similar affective polarities, and semantically opposing events have opposing affective

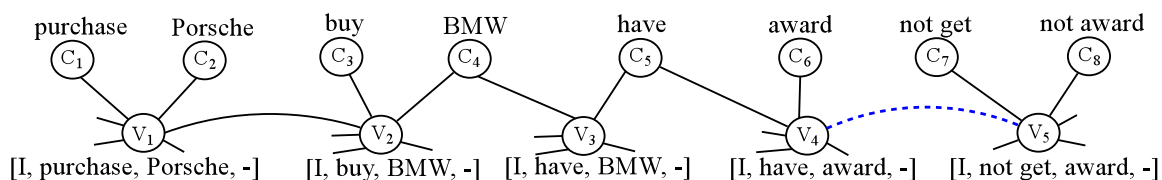


Figure 5.1: Semantic Relations Graph

polarities. Event-component relations are used by the learner to infer that the polarity of an event is related to the polarity of its individual components.

Although existing affective resources often fail to recognize many affective events, they do well at recognizing events that contain explicit emotions or strong positive/negative terms (e.g., “I had fun” or “the experience was a disaster”). Therefore, I take advantage of previously developed affective lexicons and models to provide initial polarity values for each node as noisy supervision for the SCG model.

The basic flow of the Semantic Consistency Graph model has 3 steps. First, the model builds a graph containing event and component nodes and edges representing semantic relations. Second, the model obtains initial polarities for events and components using existing sentiment analysis tools. Finally, an iterative learning algorithm is used to infer the polarities of events by optimizing the semantic consistency in the graph. The following sections present the technical details of these steps.

5.2.2 Constructing the Semantic Relations Graph

To recognize affective events, I design a semantic relations graph, which is denoted as $G = (\mathcal{V}, \mathcal{E})$ where \mathcal{V} is a set of nodes, which consists of event nodes (v_i) and component nodes (c_k). The event nodes correspond to the events extracted from the personal story corpus. The component nodes are created by decomposing each event frame representation into its parts: a predicate and up to 3 arguments. I do not create components for pronouns because they usually do not carry any polarity evidence. If a predicate is negated, then the negation is also attached to all of the event’s arguments. This simple heuristic rule considers cases where the negation of a predicate changes not only how the predicate contributes to the overall polarity of an event but also other arguments in the event. For example, the event $\langle I, \text{not get}, \text{award}, - \rangle$ will yield two component nodes: “not get” and “not award”. This strategy for handling negation can be overkill because the negation usually only applies to one part of an event. However, determining the best scope for the negation is challenging (e.g., “not have beer” is roughly equivalent to “have no beer” and for the SCG model “no beer” is more useful semantically than “not have”). In future work, more sophisticated methods need to be explored to handle the negation scope problem.

The event and component nodes in the semantic relations graph are linked with three

types of edges: *event-event similarity edges*, *event-event opposition edges*, and *event-component edges*. The following describes the construction of each type of edge in detail.

- Event-Event Similarity Edges:** The SCG model assumes that events with similar semantic meaning will usually have similar affective polarity (e.g., “have party” and “have celebration”). In this model, I use semantic embeddings of events to assess the similarities between events. I compute an event embedding as the average of the GloVe vectors (Pennington et al., 2014) of words contained in the event. Specifically, I use the 200-dimensional GloVe vectors, which were pretrained on 27 billion tweets. Since the blog corpus used in this research is also a form of social media text, the word vectors pretrained on tweets are expected to be well-suited for obtain embeddings of events extracted from the blog data. For each event node i , I first compute the similarities between i and all other events using cosine similarity of their event embeddings, then create an edge between i and each of its five most similar events. The edge weight W_{ij}^{sim} between nodes i and j is set to be the cosine similarity of their embedding vectors.
- Event-Event Opposition Edges:** The model also assumes that events with opposite meanings typically have opposite polarities (e.g., “we win” and “we did not win”). To construct opposition edges, I first identify events with a negated predicate, and refer to non-negated events as “affirmative” events. For each negated event i , I remove the negator and obtain its affirmative form \hat{i} , and compute the embedding of the affirmative form \hat{i} as the average of the embedding vectors of its words. Then, I compute the cosine similarities between the affirmative form \hat{i} and all affirmative events and select the 10 most similar affirmative events as opposition neighbors of the event i . The opposition edge weight W_{ij}^{opp} between nodes i and j is set to be the cosine similarity of embedding vectors of the affirmative form \hat{i} and its opposition neighbor j .
- Event-Component Edges:** Many event expressions refer to the same or just slightly different activities (e.g., ⟨I, have, birthday party, -⟩ and ⟨I, attend, birthday party, -⟩). Many of these events have the same affective polarity because they share the same components. Therefore, I hypothesize that learning the affective polarity of

individual component concepts could help to generalize beyond specific event frame representations. For example, if “birthday party” has positive polarity, then events mentioning a birthday party will often have positive polarity too. Of course, many events include words that have different affective polarities. However, if we link an event node with nodes for all of its components, then all of this information can be taken into account during the learning process. To explore this idea, I create edges between event node i and all of the nodes corresponding to its components. For example, the event $\langle \text{phone, fall, -, in toilet} \rangle$ will be connected with 3 component nodes: “phone”, “fall”, and “in toilet”. The edge weight between event i and component k is set to be $W_{ik}^{cmp} = 1$.

5.2.3 Learning by Optimizing Semantic Consistency

After building the semantic relations graph, the SCG model uses an iterative learning algorithm to infer the polarity of events by optimizing the semantic consistency in the graph. In this section, I present the details of the iterative learning algorithm. First, I will show how each node in the graph is initialized with a (noisy) polarity vector using existing sentiment resources. Then, I will present details of the semantic consistency measures, the objective function, and the learning algorithm. Finally, I will present a method to improve the polarity initialization for component nodes.

5.2.3.1 Initialization

In the SCG model, I associate each node with a *polarity vector* that represents a distribution over 3 polarity values $\langle \text{POSITIVE, NEUTRAL, NEGATIVE} \rangle$ for the associated event or component. This is the polarity vector to be inferred during the learning process, and the sum of the vector should be one. In addition, I also assign each event node and component node with a *initial polarity vector*, which contains polarity values over 3 class and is computed using external sentiment analysis resources. Intuitively, the idea is to initialize the model with noisy supervision, which the learner uses in combination with the semantic relations, graph structure, and optimization function to infer the correct polarity for each node. For polarity initialization, I experimented with a variety of affective lexicons and classification models and found that a method that combined the MPQA lexicon (Wilson et al., 2005) with an aggregated contextual classifier performed best. This

combination method is called “Combo”, and is described in the Evaluation section (Section 5.3).

5.2.3.2 Semantic Consistency Metrics

The SCG model infers polarity values by optimizing the semantic consistency in the graph (i.e., minimizing the inconsistency in the graph). Specifically, the inconsistency in the graph is measured as the overall inconsistency across all three types of semantic relations and between estimated values and their initial polarities. In this model, I use KL-divergence to measure the inconsistency between polarity vectors. The semantic inconsistency terms in this model are described in detail below.

- **Inconsistency between Estimated and Initial Polarities:** In the SCG model, each node (event and component) is assigned with an initial polarity vector. The model assumes that most (at least the majority) of initial polarity vectors are accurate, and encourages the estimated values to be similar to the initial vectors. Therefore, the learning algorithm tries to minimize the inconsistency between the estimated values and initial vectors. Specifically, the initial polarity vector for event i is denoted as v_i^0 , and the initial vector for component k is denoted as c_k^0 . Correspondingly, the estimated polarity vectors for event and component are denoted as v_i and c_k . Using KL-divergence, the inconsistency between v_i and v_i^0 is computed as:

$$D(v_i || v_i^0) = \sum_l v_i(l) \log \frac{v_i(l)}{v_i^0(l)}$$

where L is the set of polarity labels, and $v_i(l)$ denotes the score of polarity label l for event i . Therefore, the inconsistency for all events is computed as $\sum_{i=1}^n D(v_i || v_i^0)$ where n is the number of event nodes. Similarly, the inconsistency between the estimated polarity vectors for component nodes and their initial polarity vectors is measured as $\sum_{k=1}^m D(c_k || c_k^0)$ where m is the number of component nodes. During the learning process, the initial polarity values (i.e., v^0 and c^0) never change and serve as an anchor to prevent thrashing.

- **Inconsistency between Semantically Similar Events:** The SCG model assumes that two connected semantically similar events should have similar polarity vectors. Therefore, it aims to minimize the inconsistency between the polarity vectors of two sim-

ilar events. Specifically, for similar event pairs i and j , their inconsistency is measured as the weighted difference between their polarity vectors: $W_{ij}D(\mathbf{v}_i||\mathbf{v}_j)$. For all semantically similar pairs (i, j) , the inconsistency is computed as the sum of the inconsistencies over all pairs, i.e., $\sum_{(i,j)} W_{ij}^{sim} D(\mathbf{v}_i||\mathbf{v}_j)$.

- **Inconsistency between Semantically Opposite Events:** The model also assumes that semantically opposite events should have opposite polarity values, i.e., for opposing event pairs i and j , the positive (negative) value in i 's polarity vector should be similar to the negative (positive) value in j 's vector. To directly compute the inconsistency between two opposing event pairs, I use the *exchange matrix* $\mathbf{H} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ to switch the positive and negative values of a polarity vector. The indices of \mathbf{H} represent: 0(positive), 1(neutral), 2(negative). Therefore, the inconsistency for all opposing event pairs i and j is computed as $\sum_{(i,j)} W_{ij}^{opp} D(\mathbf{v}_i||\mathbf{v}_j\mathbf{H})$,
- **Inconsistency between Events and Their Components:** The model assumes that the polarity of most events originates from their components, and encourages events and their components to have similar polarity vectors. Therefore, the learning algorithm will also try to minimize the inconsistency between an event and each of its components. Since KL-divergence is asymmetric, the inconsistency between an event i and component k needs to be decomposed into two parts: $\sum_{(i,k)} W_{ik}^{cmp} D(\mathbf{v}_i||\mathbf{c}_k)$ and $\sum_{(k,i)} W_{ki}^{cmp'} D(\mathbf{c}_k||\mathbf{v}_i)$ where $W_{ki}^{cmp'}$ is the transpose of the event-component edge weight matrix W_{ik}^{cmp} . This maintains the symmetric property of the final objective, which allows us to directly derive closed form update equations.

5.2.3.3 Weight Normalization

When I create edges between events based on semantic similarity and opposition, for each event, I require to select a fixed and equal number of most similar (or opposite) as its neighbors. Even with this restriction, some nodes in the graph are highly connected but others are not. This is because one event can be selected many times by other events (which can be much more than the fixed number). For example, given an event i , I first select, for example, 5 most similar events and connect them to event i . However, later the event i can be selected as one of the most similar events by other events (say 1000 other events). Then, event i will totally have a degree of 1005, while some other events may just

have degrees of 5. One potential issue is that the high degree nodes may have much more influence than other nodes. To account for this, I normalize the semantic similarity weight matrix as

$$\tilde{\mathbf{W}}^{sim} = \mathbf{A}^{-\frac{1}{2}} \mathbf{W}^{sim} \mathbf{A}^{-\frac{1}{2}}$$

where \mathbf{A} is a diagonal matrix and $A_{ii} = \sum_{j=1}^n W_{ij}^{sim}$. Specifically, each entry in the new weight matrix is computed as:

$$\tilde{W}_{ij}^{sim} = \frac{W_{ij}^{sim}}{\sqrt{A_{ii}} \sqrt{A_{jj}}}$$

I similarly normalize the semantic opposition weight matrix \mathbf{W}^{opp} to obtain $\tilde{\mathbf{W}}^{opp}$.

For the event-component edges, different events may link to different numbers of components, and vice versa. For example, some events may only have 2 components and they will be connected to the 2 components, while others may have 4 components. In addition, some components (e.g., “have”) can be connected to a lot of events in which they appear, and these components will have high degrees in the graph. To account for this, I also normalize the weights for event-component edges. To normalize the weights \mathbf{W}^{cmp} of edges from events to components, I compute the new weight matrix $\tilde{\mathbf{W}}^{cmp}$ by performing row normalization on the original matrix, i.e.,

$$\tilde{W}_{ik}^{cmp} = \frac{W_{ik}^{cmp}}{D_i}$$

where $D_i = \sum_k W_{ik}$. Similarly, I also compute the normalized weight matrix $\tilde{\mathbf{W}}^{cmp'}$ for edges from components to events by performing row normalization on the original matrix $\mathbf{W}^{cmp'}$, which is the transpose of \mathbf{W}^{cmp} .

5.2.3.4 The Objective and Update Functions

The final Semantic Consistency Graph model uses normalized edge weights and incorporates all of the previously described inconsistency measures into a single objective function, which is shown in Equation 5.1.

$$\begin{aligned} J_{sc} = & \beta \sum_{i=1}^n D(\mathbf{v}_i || \mathbf{v}_i^0) + \sum_{(i,j)} \tilde{W}_{ij}^{sim} D(\mathbf{v}_i || \mathbf{v}_j) + \sum_{(i,j)} \tilde{W}_{ij}^{opp} D(\mathbf{v}_i || \mathbf{v}_j \mathbf{H}) + \gamma \sum_{(i,k)} \tilde{W}_{ik}^{cmp} D(\mathbf{v}_i || \mathbf{c}_k) \\ & + \gamma \sum_{(k,i)} \tilde{W}_{ki}^{cmp'} D(\mathbf{c}_k || \mathbf{v}_i) + \eta \sum_{k=1}^m D(\mathbf{c}_k || \mathbf{c}_k^0) \quad (5.1) \end{aligned}$$

This objective computes the overall inconsistency in the graph, and the goal is to minimize the objective to obtain the best polarity estimates. The n and m are the numbers of event and components nodes, and (i, j) denotes connected node pairs for each type of relations. The hyperparameters control the relative importance of each corresponding term. In the experiments presented in Section 5.3, the full model uses the following values: $\beta = 0.6$, $\gamma = 0.8$, $\eta = 0.1$, which were selected based on the performance on the development data.

Since KL-divergence is convex, the objective in Equation 5.1 is convex when the parameters are non-negative. This guarantees that the model will converge at a global minimum. I designed an iterative algorithm that alternately updates v and c . Let v_i^t and c_k^t denote polarity vectors for event i and component k at iteration t . The learning algorithm first optimizes the objective over v_i given v_i^t and c_k^t . By computing the derivative for v_i^{t+1} , the update for v_i^{t+1} is shown in Equation 5.2. The detailed derivation is shown in Appendix C.

$$v_i^{t+1} \propto \exp \frac{1}{O_i} \left(\beta \log v_i^0 + \sum_j \tilde{W}_{ij}^{sim} \log v_j^t + \sum_j \tilde{W}_{ij}^{opp} \log v_j^t \mathbf{H} + \gamma \sum_k \tilde{W}_{ik}^{cmp} \log c_k^t \right) \quad (5.2)$$

where $O_i = \beta + \sum_j \tilde{W}_{ij}^{sim} + \sum_j \tilde{W}_{ij}^{opp} + \gamma \sum_k \tilde{W}_{ik}^{cmp}$.

Given v_i^{t+1} , the update equation for c_k^{t+1} can be obtained by computing the derivative for c_k^{t+1} . The update equation is shown in Equation 5.3, and the derivation is described in Appendix C.

$$c_k^{t+1} \propto \exp \frac{\eta \log c_k^0 + \gamma \sum_i \tilde{W}_{ki}^{cmp'} \log v_i^{t+1}}{\eta + \gamma \sum_i \tilde{W}_{ki}^{cmp'}} \quad (5.3)$$

The learning algorithm is shown in Algorithm 2, which iteratively updates the polarity vectors on event nodes and component nodes until the polarity values on event nodes converge. In the final experiments presented in Section 5.3, the learning process converged after 52 iterations. When the learning is finished, each event i is associated with an estimated polarity vector with 3 polarity values, and I infer its polarity label to be the polarity receiving the highest score, i.e., $\arg \max_l v_i(l)$. As I will discuss in the evaluation section (Section 5.3), alternately, a threshold can be applied to only assign polarity labels to events when a high percentage of probability mass is associated with one polarity.

5.2.3.5 Improved Component Initialization

Instead of initializing component nodes directly from sentiment analysis resources, I hypothesized that we could improve the initial polarity values of components through

Algorithm 2 Iterative Learning Algorithm

- 1: **Input:** $W^{sim}, W^{opp}, W^{cmp}, v^0, c^0$
 - 2: **Output:** $v \in \mathcal{R}^{n \times |L|}$
 - 3: **while** v has not converged **do**
 - 4: Update v^t using Equation 5.2
 - 5: Update c^t using Equation 5.3
 - 6: **end while**
 - 7: **return** v^t
-

an independent learning process that exploits semantic similarities among components. To explore this idea, I create a separate component graph that contains only component nodes with edges that connect each component node to its 5 most similar component nodes. The similarity between components is estimated as the cosine similarity between their embedding vectors, and a component embedding is computed as the average of embeddings for the words in the component. I also used the GloVe vectors and set the edge weight U_{ij} between component i and j to be the cosine similarity of their embeddings. In this component graph, the semantic inconsistency comes from the inconsistency between a component’s initial value and its estimated value, and between the connected components. The final inconsistency (J_{cmp}) is shown in Equation 5.4 where m is the number of component nodes.

$$J_{cmp} = \sum_{(i,j)} \tilde{U}_{ij} D(c_i || c_j) + \sum_{i=0}^{m_l} D(c_i || c_i^s) + \mu \sum_{i=0}^m D(c_i || c_i^0) \quad (5.4)$$

The first term in Equation 5.4 measures the inconsistency between two semantically similar components. The second term measures inconsistency between the estimates and initial polarities (c^s) computed from MPQA Lexicon. Since not all component words are contained in MPQA, the m_l denotes the components found in MPQA. The third term measures inconsistency between the estimated values and initial polarities (c^0) assigned by a contextual classifier (NRC^{AvgS}), which is described in Section 5.3. I use both MPQA and NRC^{AvgS} to initialize this model because the MPQA lexicon has high precision but low coverage, while the classifier has better coverage but lower precision. I hypothesize that using both of these methods as noisy supervision can potentially achieve better performance. Since NRC^{AvgS} is not very accurate, it is associated with a penalizing parameter μ to reduce its importance and influence in the model. Given the objective, I derive the update

function for variable c_k by computing its derivative, and iteratively updating the polarity values for 100 iterations. In the experiments, I used $\mu=0.1$ based on the performance on the development set. The inferred polarity vectors are then used as the “initial” polarity vectors for the component nodes in the full SCG model. This separate learning process for component nodes slightly improved the overall evaluation results.

5.3 Evaluation

This section presents experiments for evaluating the performance of the Semantic Consistency Graph model. As comparison, this section first investigates the performance of three types of baseline methods: affective lexicon-based methods, classification models, and a combination method based on a lexicon and a contextual classification model. Then, this section demonstrates the value of the three types of semantic relations by incrementally adding them into the model. Experiments were conducted on a set of randomly sampled events with manual polarity annotations created in Section 3.3. Model parameters used in experiments were tuned on the development set, and all results are reported on the test set.

5.3.1 Performance of Affective Lexicons and Learning Models

In this section, I investigate the performance of methods based on previous sentiment analysis resources. First, I evaluate the effectiveness of methods based on affective lexicons, and the performance of two types of classification models: event expression classifiers and contextual classifiers. Finally, I create a combination method using both a lexicon and a contextual classification model. The evaluation details for these methods are presented below.

I evaluated the performance of five existing affective lexicons that are **MPQA** (Wilson et al., 2005), **SentiWordNet3.0 (SentiWN)** (Baccianella et al., 2010), **+/-EffectWordNet (+/-EffectWN)** (Choi and Wiebe, 2014), **ConnotationWordNet (ConnoWN)** (Kang et al., 2014), and **Connotation Frames** (Rashkin et al., 2016) for which I evaluated both the effect on subject (**ConnoFrameS**) and the effect on object (**ConnoFrameO**) lexical information. Since the event structures contain multiple words, I computed the polarity score for an event as the average over the scores of its words. Most of these lexicons assign polarity scores over

a range of values to capture the strength of positive or negative polarity. To explore the best way to use each lexicon, I defined a threshold λ for each lexicon. For lexicons with polarity values ranging from $[-1,+1]$, I assigned events with a score $> \lambda$ as positive, $< -\lambda$ as negative, and an absolute value $|\text{score}| \leq \lambda$ as neutral. For lexicons with polarity values ranging from $[0,+1]$, I assigned events with a score between $[0.5-\lambda, 0.5+\lambda]$ as neutral, $< 0.5-\lambda$ as negative, and $> 0.5+\lambda$ as positive. I found that the following values achieved the best F1 scores on the development data and were therefore used throughout our experiments: $\lambda=0$ for MPQA, $\lambda=0.25$ for ConnoFrameS, $\lambda=0.3$ for ConnoFrameO, $\lambda=0.4$ for ConnoWN, $\lambda=0.5$ for SentiWN, and $\lambda=0.6$ for +/-EffectWN.

The experimental results for lexicons and learning models are presented in Tables 5.1 and 5.2. Table 5.1 shows the F1 scores for the positive (POS), negative (NEG), and neutral (NEU) polarities, and the macro-averaged F1 score across all three polarities, and Table 5.2 presents precision (Pre) and recall (Rec) scores for each polarity class.

The top portion of Tables 5.1 and 5.2 shows the performance of affective lexicons. These results show that the MPQA lexicon performs the best, achieving the best macro F1 score and the highest precision for each polarity class, among these lexicon-based methods.

For learning-based methods, I first evaluated several *event expression classifiers* by ap-

Table 5.1: F1 Scores for Lexicons, Event Expression Classifiers, and Contextual Models

Method	POS	NEG	NEU	AVG
Affective Lexicons				
ConnoWN	26.3	9.8	64.1	33.4
ConnoFrameS	32.6	21.0	64.8	39.5
ConnoFrameO	29.8	22.5	70.7	41.0
+/-EffectWN	36.3	36.7	55.3	42.8
SentiWN	33.5	27.3	73.9	44.9
MPQA	57.8	54.9	80.1	64.3
Event Expression Classifiers				
LR ^{BOW}	25.6	16.1	78.2	40.0
StanfordSA	37.5	12.4	77.7	42.6
LR ^{Embed}	50.8	44.9	79.7	58.5
NRC	58.6	55.9	79.6	64.7
Contextual Models				
ECC	28.1	46.1	65.9	46.7
NRC ^{AvgS}	51.2	52.0	70.7	58.0
Combo	60.7	58.3	79.9	66.3

Table 5.2: Precision and Recall for Lexicons, Event Expression Classifiers, and Contextual Models

Method	POS		NEG		NEU	
	Pre	Rec	Pre	Rec	Pre	Rec
Affective Lexicons						
ConnoWN	23.9	29.3	35.7	5.6	59.5	69.4
ConnoFrameS	28.1	38.9	54.8	13.0	62.0	67.8
ConnoFrameO	32.2	27.8	55.6	14.1	63.5	79.7
+/-EffectWN	26.2	59.1	47.3	29.9	66.7	47.2
SentiWN	36.0	31.3	50.8	18.6	67.2	82.1
MPQA	64.2	52.5	60.5	50.3	76.3	84.3
Event Expression Classifiers						
LR ^{BOW}	61.5	16.2	39.1	10.2	66.2	95.5
StanfordSA	65.8	26.3	40.6	7.3	66.1	94.1
LR ^{Embed}	63.2	42.4	56.4	37.3	73.1	87.7
NRC	75.4	48.0	53.3	58.8	76.4	83.0
Contextual Models						
ECCG	68.6	17.7	34.9	67.8	66.9	64.8
NRC ^{AvgS}	55.6	47.5	40.8	71.8	77.9	64.8
Combo	67.7	55.1	56.3	60.5	78.4	81.4

plying them directly to the sequence of words in an event structure. I approximately replicated the NRC-Canada sentiment classifier (**NRC**) (Mohammad et al., 2013), and trained the classifier using the SemEval 2014 Task 9 tweet data. The details of this re-implementation are described in Section 4.2.2. I also evaluated the Stanford sentiment analysis (**StanfordSA**) system (Socher et al., 2013), which is a neural network model. In addition, I trained two logistic regression classifiers on the development data using the Scikit-learn toolkit (Pedregosa et al., 2011). One classifier (**LR^{BOW}**) uses bag of words features for all words in an event. A second classifier (**LR^{Embed}**) uses word embedding features, which is computed as the average of the word embedding vectors over the words in the event representation. The word embeddings are the 200-dimensional GloVe vectors pretrained on 27B tweets. The middle of Tables 5.1 and 5.2 shows the performance of the event expression classifiers. The results show that the NRC classifier achieved the best macro F1 score among these systems, which demonstrates that sentiment analysis classifiers trained on tweets data can recognize 48% positive events and 58% negative events with 75% and 53% precision, respectively.

In addition, in this research, I hypothesized that the affect expressed in the contexts

where an event is mentioned can be useful for predicting the affective polarity of the event. Therefore, I also evaluated two types of *contextual models*, which exploit the contexts surrounding an event. For each event, I applied the NRC classifier to every sentence in which it occurs and produced a distribution of polarity values across the sentences. I call this method NRC^{AvgS} . I also evaluated the previous Event Context Graph (ECG) model described in Chapter 4 on this new set of randomly sampled events. I applied it to the same blog texts data set of nearly 1.4 million blog posts. The ECG model produces polarity values ranging from $[-1, +1]$, so I tuned a λ parameter on the development data as I did for the lexicons, and obtained the best value: $\lambda=0.15$. Table 5.1 shows that NRC^{AvgS} was the best performing contextual model.

MPQA was the best lexicon, and NRC^{AvgS} was the best contextual model, and I hypothesized that combining these complementary methods might perform even better. Therefore, I created a **Combo** system that linearly combines the predictions of both models. For an event e , I compute its polarity vector as:

$$\alpha * \text{PolarityVector}_{\text{NRC}^{\text{AvgS}}}(e) + (1 - \alpha) * \text{PolarityVector}_{\text{MPQA}}(e)$$

where $\text{PolarityVector}_{\text{NRC}^{\text{AvgS}}}(e)$ denotes the polarity vector of event e , which is computed using the NRC^{AvgS} system and consists of polarity values over 3 classes (i.e., positive, negative, and neutral). $\text{PolarityVector}_{\text{MPQA}}(e)$ is the polarity vector obtained using MPQA lexicon. The last row of Table 5.1 shows the results for this **Combo** method. The parameter α is set to be 0.7 based on the development set. Results show that this Combo method achieved the highest F1 score, which demonstrates the effectiveness of combining sentiment lexicons and contextual models, and that these two types of methods are complementary to each other.

5.3.2 Performance of the Semantic Consistency Graph Model

In this section, I evaluate the performance of the Semantic Consistency Graph (SCG) model, and analyze the effectiveness of each type of semantic relation for improving performance.

Table 5.3 shows the results for the SCG model alongside the best system (**Combo**) that utilized existing methods, for comparison. I initialized the polarity vectors of the event

Table 5.3: Results for Semantic Consistency Graph (SCG) Model

Method	POS	NEG	NEU	Average		
	F1	F1	F1	Precision	Recall	F1
Combo	60.7	58.3	79.9	67.5	65.6	66.3
SCG+sim	58.6	62.9	82.3	72.6	65.7	67.9
+opp	59.9	63.8	83.4	75.0	65.8	69.0
+cmp	63.7	66.7	83.7	75.2	68.9	71.4

nodes in the SCG model using the **Combo** method, which produces a distribution over the 3 polarity values for each event.

The **SCG+sim** row first shows results for the SCG model using only the semantic similarity edges, which substantially improves precision (+5%) over the Combo baseline. The **+opp** row shows results for adding the semantic opposition edges as well, which further improves precision to 75% while maintaining the same level of recall. The **+cmp** row shows results for the full model, which also includes component nodes connected to corresponding events. These shared component relations improve recall from 65.8% to 68.9% without any loss of precision. Overall, the full semantic consistency model achieved both higher recall (65.6% \rightarrow 68.9%) and higher precision (67.5% \rightarrow 75.2%) compared to the best results achieved with previous methods. The macro-averaged F1 score improved from 66.3% to 71.4%, which is statistically significant at $p < 0.01$ based on the paired bootstrap test (Berg-Kirkpatrick et al., 2012).

5.4 Analysis

In this section, I first analyze the behavior of the Semantic Consistency Graph (SCG) model using the gold human annotations, and then examine both the quality and quantity of affective events identified by the SCG model.

5.4.1 Error Analysis

First, I identified all events in the test set whose initial polarity, which was estimated by the Combo model, was changed by the SCG model. Table 5.4 shows the number of events whose polarity labels (i.e., positive (POS), negative (NEG), and neutral (NEU)) were changed. The first column shows the polarity labels that were initially assigned by the Combo system and then labeled by the SCG model. #Total denotes the total number of

Table 5.4: Polarity Changes between Combo and SCG models

Combo → SCG	#Total	#Correct	Accuracy
POS → NEU	24	19	79%
NEU → NEG	18	13	72%
NEG → NEU	45	32	71%
POS → NEG	4	2	50%
NEG → POS	8	3	38%
NEU → POS	3	1	33%

events whose labels are changed. #Correct denotes the totally correct number of events whose polarity labels were changed. For example, the first row (i.e., POS → NEU) shows that there are a total of 24 events that were initially predicted as positive (POS) by the Combo method and labeled as neutral (NEU) by the SCG model, among which 19 events are correctly changed, resulting in an accuracy of 79%. This analysis shows that the most frequent changes were from positive or negative polarity to neutral, and from neutral to negative.

Table 5.5 shows the precision and recall differences between Combo and SCG models for each polarity. The large shifts from positive/negative to neutral correspond to the precision gains, and the shifts from neutral to negative correspond to the increased recall for negative polarity.

In addition, I also looked at some concrete examples whose labels were changed by the SCG model. Table 5.6 shows some correct and incorrect examples of events whose polarity changed from X to Y. For example, the top portion of Table 5.6 shows examples whose polarities were changed from positive to neutral. The SCG model seems to have learned that certain predicates (verbs) are typically neutral, such as “open” and “want”. I also observe that many of its errors involve negated terms, suggesting that more sophisticated negation handling may be needed.

Table 5.5: Precision and Recall Breakdowns for Combo and SCG Model

Method	POS		NEG		NEU	
	Precision	Recall	Precision	Recall	Precision	Recall
Combo	67.7	55.1	56.3	60.5	78.4	81.4
SCG	75.7	55.1	70.4	63.3	79.3	88.5

Table 5.6: Correct and Incorrect Examples

POSITIVE → NEUTRAL	
Correct Examples:	⟨ I, open, my email, - ⟩
⟨ box, be, open, - ⟩	⟨ my friend, start, work, - ⟩
⟨ I, want, photo, - ⟩	⟨ I, want, bag, - ⟩
Incorrect Examples:	⟨ my family, stay, with me, - ⟩
⟨ I, win, class, -, - ⟩	⟨ band, rock, -, - ⟩
NEUTRAL → NEGATIVE	
Correct Examples:	⟨ food, not be, tasty, - ⟩
⟨ I, break, heart, - ⟩	⟨ friend, disappoint, me ⟩
⟨ I, be, bummed, - ⟩	⟨ I, start, sniffle, - ⟩
⟨ tear, pour, -, from eye ⟩	⟨ none, be, -, for me ⟩
Incorrect Examples:	
⟨ we, see, cave, - ⟩	⟨ we, steal, glance, - ⟩
NEGATIVE → NEUTRAL	
Correct Examples:	⟨ feeling, go, -, through me ⟩
⟨ I, feel, -, about stuff ⟩	⟨ I, need, bowl, - ⟩
⟨ I, call to work, -, - ⟩	⟨ answer, not be, one, - ⟩
Incorrect Examples:	⟨ my memory, not serve, me, - ⟩
⟨ house phone, not work, -, - ⟩	⟨ I, not function, -, at work ⟩

5.4.2 Quality and Quantity of the Learned Affective Events

A goal of this research is to produce an affective event collection that can be used by the NLP community as knowledge of affective events. Toward this end, I created lexicons of varying sizes by selecting events that were assigned a positive or negative polarity with value $\geq \tau$ in the polarity vector, to effect recall/precision trade-offs. Table 5.7 shows the precision and recall on the test set when assigning polarity labels using different thresholds, and also the total number of affective events extracted from the corpus for the corresponding thresholds. The bottom row (*max*) shows the lexicon produced by assigning every event the polarity that has the highest value with no additional threshold. The bottom row shows that the complete lexicon has over 175K affective events with precision $>70\%$. Setting $\tau=0.5$ still produces 111K events with $>80\%$ precision for negative and $>90\%$ for positive events. Increasing the threshold to 0.6 reduces the lexicon to approximate $>69,000$ affective events with $>93\%$ precision for both polarities. Examples of the automatically learned affective events are presented in Appendix D.

In addition, I notice that the SCG model produced more negative than positive events. This phenomenon is consistent with the results of the initialization method (i.e., the Combo

Table 5.7: Quality and Size of Affective Event Collections Extracted with Different Thresholds

τ	POSITIVE		NEGATIVE		#AffectiveEvents		
	Precision	Recall	Precision	Recall	#pos	#neg	#total
0.7	100	18.7	93.7	16.9	19031	18947	37978
0.6	96.9	31.8	93.4	32.2	30584	38523	69107
0.5	90.1	41.4	80.2	45.8	48594	62998	111592
max	75.7	55.1	70.4	63.3	82398	92743	175141

method), which are obtained by comparing gold polarities in the test set (i.e., the Combo results in Table 5.5). Therefore, I believe this phenomenon is influenced by the initialization method.

Overall, this analysis shows that the SCG model can be used to automatically generate large, high-quality collections of affective events. The learned affective events with affective polarities are called **Affective Event Knowledge Base** (AffectEventKB), and are released and freely available for the research community.

5.5 Chapter Summary

This chapter presents the technical details of the Semantic Consistency Graph (SCG) model and experiments for evaluating the model. This chapter also provides an analysis of both the quality and quantity of affective events identified by the SCG model. The main content of this chapter is summarized below.

- This chapter presents a Semantic Consistency Graph model to identify affective events from personal blog posts using semantic relations between events. First, the model creates a Semantic Relations Graph consisting of event nodes and component nodes, which are connected with *event-event similarity* edges, *event-event opposition* edges, and *event-component* edges. Each node in the graph is assigned with an initial polarity value using sentiment analysis resources. Then, the model uses an iterative learning algorithm to infer correct polarities of events by optimizing the semantic consistency in the graph.
- This chapter describes experiments for evaluating the SCG model and comparing it to three types of baseline methods: affective lexicon-based methods, event expression classifiers, contextual learning models, and a combination method. Experimen-

tal results show that the SCG model significantly outperforms these other methods. Further analysis also demonstrates that the three types of semantic relations all contribute to improving the performance of identifying affective events.

- This chapter analyzes the behavior of the SCG model, and also examines both the quality and quantity of affective events extracted with different confidence thresholds. The analysis shows that the model gains precision by correctly changing positive or negative labels to neutral. Overall, the SCG model can identify a large set of affective events with high precision. Specifically, the model identified over 110,000 affective events with over 90% precision for positive events and over 80% precision for negative events.

CHAPTER 6

HUMAN NEEDS CATEGORIZATION OF AFFECTIVE EVENTS USING LABELED AND UNLABELED DATA

The research described in Chapters 4 and 5 focuses on learning affective polarity of events. However, when we discuss an event, we not only understand its affective polarity but also the reason why the event is beneficial or detrimental. For example, when someone says that “I broke my leg” we not only know that the person was impacted negatively but also understand that the person has a health problem. Conversely, for the event “I got a job”, we understand not only that the experiencer would typically feel happy but also know that the reason is that the person has a financial income.

This chapter aims to understand the reason for events being affective, i.e., to answer why an event is positive or negative. Events can impact people in many ways, and understanding *why* an event is beneficial (positive) or detrimental (negative) is a fundamental aspect of language understanding and narrative text comprehension. Additionally, many applications can potentially benefit from understanding affective events, including text summarization, conversational dialogue processing, and mental health therapy or counseling systems. As an illustration, a mental health therapy system can benefit from understanding why someone is in a negative state. If the triggering event for depression is “*I broke my leg*”, then the reason is related to the person’s Health, but if the triggering event is “*I broke up with my girlfriend*” then the reason is based on Social relationships.

In this chapter, I hypothesize that the polarity of affective events can often be attributed to a relatively small set of *human need* categories. My research is motivated by theories in psychology that explain people’s motivations, desires, and overall well-being in terms of categories associated with basic human needs, such as Maslow’s Hierarchy of Needs (Maslow et al., 1970) and Fundamental Human Needs (Max-Neef et al., 1991). Drawing

upon these works, I propose that the polarity of affective events often arises from 7 types of human needs: PHYSIOLOGICAL, HEALTH, LEISURE, SOCIAL, FINANCIAL, COGNITION, and FREEDOM. For example, “*I broke my arm*” has negative polarity because it negatively impacts one’s Health, “*I got fired*” is negative because it negatively impacts one’s Finances, and “*I am confused*” is negative because it reflects a problem related to Cognition.

To validate the above hypothesis, I conducted a manual annotation study on the affective events that were previously manually identified as positive or negative. In the study, annotators were asked to assign a human need category label to each affective event. The annotation results show that the majority of the affective events can be manually categorized into the 7 human need categories with good inter-annotator agreements (Cohen’s $\kappa \geq .65$). The resulting affective events with human need category labels are used as evaluation data for automatic classification methods, and will be further described in Section 6.1.

To automatically categorize an affective event by the human need categories, I designed several types of supervised and semi-supervised classification models that learn from both labeled and unlabeled data. First, I built supervised learning models to exploit lexical and embedding features for the words in event expressions, as well as models that learn from the sentence contexts surrounding event mentions. Then, I explored a self-training method to exploit both labeled and unlabeled data for training. I also designed a co-training model that takes advantage of a large set of unlabeled events and two different types of classifiers in an iterative learning process: event expression classifiers only rely on the words in an event expression, and event context classifiers use features from the contexts surrounding the instances of event expressions. Experimental results show that the co-training model substantially improves human need categorization performance compared to classifiers trained only with labeled data, yielding gains in both precision and recall.

In the rest of this chapter, I first present the definition and examples for each human need category, and a manual annotation study. Then, I describe the methods for categorizing affective events, including supervised classifiers, and semi-supervised models. Finally, I evaluate these classification methods on gold annotations, and summarize the main content of this chapter.

6.1 Human Need Categories and Annotations

This section presents definitions and examples of human need categories, and describes the manual annotation process for obtaining gold human need labels for affective events.

6.1.1 Human Need Categories

When we comprehend a narrative story, we understand not only the affective polarity of the events in it, but also the reason why the events are affective (positive or negative). I hypothesized that affective polarity of events arises from the satisfaction or violation of basic human needs. Psychologists have developed theories that explain people's motivations and personalities, and communities' strength and weakness in terms of categories associated with basic human needs, such as Maslow's Hierarchy of Needs (Maslow et al., 1970) and Fundamental Human Needs (Max-Neef et al., 1991). In this research, I developed a set of 7 human need categories by selecting and separating some categories from these two theories. In addition, I defined 2 other classes for categorizing affective events that cannot be classified into the 7 human need categories. In the following, I will first describe how my 7 human need categories relate to the two theories. Then, I will present the definition and examples for each of the human need categories.

In Maslow's Hierarchy of Needs (MHN), human needs are categorized into: *Physiological Needs, Safety Needs, Belonging and Love Needs, Esteem Needs, Cognitive Needs, Aesthetic Needs, Self-actualization Needs, and Self-transcendence Needs* categories. In the Fundamental Human Needs (FHN) theory (Max-Neef et al., 1991), human needs are classified into: *Subsistence Needs, Protection Needs, Affection Needs, Understanding Needs, Participation Needs, Leisure Needs, Creation Needs, Identity Needs, and Freedom Needs* classes. In this dissertation, I proposed to categorize affective events into 7 human need categories: *Physiological Needs; Physical Health and Safety Needs; Leisure and Aesthetic Needs; Social, Self-Worth, and Self-Esteem Needs; Finances, Possessions, and Job Needs; Cognition and Education Needs; and Freedom of Movement and Accessibility Needs.*

Table 6.1 shows the relations between the human need categories that I assigned to affective events and the categories in Maslow's Hierarchy of Needs and Fundamental Human Needs theories. In this dissertation, the *Physiological Needs* category corresponds to the *Physiological Needs* in MHN, and has overlap with the *Subsistence Needs* in FHN.

Table 6.1: Relations between the Human Need Categories Assigned to Affective Events for this Research and Maslow’s Hierarchy of Needs (MHN) and Fundamental Human Needs (FHN). The * Denotes that the MHN or FHN Categories Have Overlap with My Category in the Same Row, and the + Denotes that the MHN or FHN Categories Are a Subset of My Corresponding Category.

Human Need Category for Affective Events	MHN Category	FHN Category
Physiological	Physiological	Subsistence*
Physical Health and Safety	Safety*	Protection*
Leisure and Aesthetic	Aesthetic ⁺	Leisure ⁺
Social, Self-Worth, and Self-Esteem	Belonging and Love ⁺ , Esteem ⁺	Affection ⁺ , Identity ⁺
Finances, Possessions, and Job		
Cognition and Education	Cognition	Understanding
Freedom of Movement and Accessibility		Freedom*

It does not include the mental health needs from the Subsistence category because most events related to mental health are sentiment or emotion expressions, which are categorized into the *Sentiments, Emotions, or Opinions* category in my research. In my data, the mental or emotion expressions are a big class, and I feel that separating them out from other physiological events is beneficial to understand the distribution of affective events in the data. The *Physical Health and Safety Needs* category in this dissertation is a set of human needs related to the health and safety of a physical body, similar to Safety in MHN and Protection in FHN. However, it does not include the emotional security needs and the need for law and order from the Safety Needs in MHN. This category also does not include the need for autonomy and social security that are included in the Protection Needs in FHN, because they were primarily ascribed to group dynamics, not individuals.

My *Leisure and Aesthetic Needs* contain both the Aesthetic Needs in MHN and the Leisure Needs in FHN. I combined them together because many events related to the Aesthetic Needs are also associated with the Leisure Needs. For example, “I watch a sunset” indicates the satisfaction of both Aesthetic Needs and Leisure Needs. In addition, I merged the Belonging and Love Needs and Esteem Needs from MHN, and the Affection and Identity Needs from FHN into the *Social, Self-Worth, and Self-Esteem Needs* since these human needs

are about social relations with others (e.g., family, and friends), social positions, and self-worth in society. My *Cognition and Education Needs* category basically corresponds to the Cognition Needs in MHN and the Understanding Needs in FHN into the *Cognition and Education Needs*, and they are all about learning knowledge, skills, or receiving education. In this dissertation, the *Freedom of Movement and Accessibility Needs* category is mainly about physical freedom, and is related to the Freedom Needs in FHN that also contain the needs for autonomy and spiritual freedom.

In addition, I created a new *Finances, Possessions, and Job Needs* category for events related to finance, valuable possessions, and jobs, which does not have a direct corresponding category in the two theories. In my research, I did not use the Self-actualization and the Self-transcendence Needs in MHN, or the Participation Needs and the Creation Needs in FHN for categorizing affective events because I found that few affective events related to these human need categories in my corpus.

In the following paragraphs, I will present the definitions and examples for each human need category studied in this dissertation.

- **Physiological Needs**

Physiological needs are the basic needs that must be satisfied to maintain the human body's basic functions. For example, we need to eat food, our body needs sleep, etc. An event in this category is affective often because it implies the satisfaction or violation of a Physiological need. For example, "I am hungry" is negative because the need to have food is not satisfied.

Specifically, Physiological needs include (1) the need to be able to breathe beneficial or pleasant air, and to avoid detrimental or unpleasant air; (2) the need to avoid hunger, to avoid unpleasant food, and to eat or obtain pleasing food; (3) the need to avoid thirst, to avoid unpleasant beverages, and to drink or obtain pleasing beverages; (4) the need to sleep, regularly and comfortably; (5) the need to maintain warmth, to not be too hot or too cold; (6) the need to have or obtain shelter (i.e., a place to live or stay) and to avoid unpleasant shelters. If an event is affective and indicates the satisfaction or violation of these needs, then it belongs to this category.

Examples

- *"I have not eaten for 2 days"* is negative because the need to have food is violated.
- *"I woke up at 2am"* is negative because the speaker violated the need of having enough sleep or sleeping soundly.
- *"I ate cake"* is positive because the need of having enjoyable food is satisfied.
- *"I bought a house"* is positive because the need of owning a shelter is satisfied.

- **Physical Health and Safety Needs**

In our daily lives, many events will harm or endanger our physical health and safety. If we experience these events, we are often affected negatively. Conversely, some other events will improve our physical health or safety conditions and affect us positively when we experience them. These events are affective because they satisfy or violate the needs to be physically healthy and safe. Affective events in this category could be related to health problems, body injuries, exercise, etc.

Examples

- *"My head hurts"* is negative because the need to be physically healthy is violated.
- *"I do exercise"* is positive because exercise is associated with good health.
- *"I was kidnapped"* is negative because the speaker is concerned about his/her physical health and safety.

- **Leisure and Aesthetic Needs**

Another type of need that most people share is the need to have fun, to be relaxed, to have leisure time. For example, people often feel happy when they experience positive events like "play games", but are affected negatively by some other events because they prevent people from having leisure time. For example, "work on holidays" is undesirable because people expect to have fun or leisure time on holidays. In addition, people need to be able to appreciate and enjoy the beauty of certain things and will typically feel relaxed, joyful, or peaceful when these needs are satisfied. Since Leisure needs and Aesthetic needs are closely related, I combine them together as a single category.

Specifically, Leisure and Aesthetic needs include: (1) the need to have entertaining or fun activities, to avoid lack of fun or entertaining activities; (2) the need to have leisure, to avoid too much work because it detracts from leisure time; (3) the need to have an enjoyable, pleasant environment; (4) the need to pursue and appreciate the beauty of nature, art, music, and other aesthetically beautiful things.

Examples

- *"I play computer games"* is positive because it describes a fun activity.
- *"I work on Christmas Day "* is negative because it is typically unenjoyable and detracts from leisure time.
- *"Room is noisy "* is negative because the environment is undesirable.
- *"I saw a rainbow"* is positive because the need to appreciate beauty is satisfied.

• Social, Self-Worth, and Self-Esteem Needs

According to Aristotle, human beings by nature are "social animals", which means that people naturally need to have close relationships with family and friends. Events indicating the satisfaction of having good family or friend relations make people feel good. Conversely, people would be impacted negatively if these needs are violated. As a member of society, people also have needs to have and improve self-worth and self-esteem, and to be respected by others.

I group these Social needs together as a single human need category. Specifically, these needs include: (1) the need to have family, to have close family relations, to avoid damaging family relations; (2) the need to have friendships, to avoid damaging friendships; (3) the need to maintain pleasant social relations, to avoid conflicts and arguments; (4) the need to maintain socially and culturally acceptable behaviors; (5) the need to realize and improve one's self-worth, to be recognized by others; (6) the need to maintain and improve self-esteem or dignity.

Examples

- *"My mom visited me"* is positive because a family relationship is maintained.
- *"I have many friends"* is positive because the friendship need is satisfied.

- *"Nobody talks to me"* is negative because social relations with others are not good.
- *"They mock me"* is negative because the speaker's self-esteem/dignity is hurt.

- **Finances, Possessions, and Job Needs**

In our daily lives, many affective events involve earning money, having well-paid jobs, and obtaining useful or valuable possessions. When these events happen to us, we would usually have positive feelings. However, if we experienced affective events like losing money, getting fired from a job, or losing some valuable possessions, we would typically be affected negatively. More specifically, this category includes: (1) the need to obtain and protect financial income; (2) the need to acquire possessions and maintain good condition of one's possessions; (3) the need to have a job and satisfying work because jobs are usually reliable sources to obtain financial income. Since possessions could be anything valuable, I define that if the possession in an event is more directly related to another type of need, I prefer to categorize the event in that human need category. For example, the event *"I bought steaks"* mainly satisfies the Physiological needs because the purpose of *"steaks"* is for eating, so it is classified into the Physiological need category rather than this category.

Examples

- *"I got a lot of money"* is positive because the need to have financial income is satisfied.
- *"I bought a computer"* is positive because the need to obtain useful tools is satisfied.
- *"I lost my watch"* is negative because the need to protect possessions is violated.
- *"I got fired"* is negative because the need to have a job has been violated.

- **Cognition and Education Needs**

Cognition and education needs motivate people to learn skills, obtain information, understand meanings, improve cognitive abilities, etc. When such needs are satisfied (e.g., *"learned new skills"*), people would often feel good; otherwise, people are

affected negatively (e.g., “I did not get the Master degree”). More specifically, this group of needs include: (1) the need to obtain skills, information, and knowledge, to receive education, and to improve one’s intelligence; (2) the need to mentally process information correctly (e.g., remembering, calculation), and to have good cognitive abilities.

Examples

- “*I learned to mow the lawn*” is positive because the speaker learned a skill.
- “*I graduated*” is positive because the need to receive education is satisfied.
- “*I overestimated the number*” is negative because the speaker did not process information correctly.

● Freedom of Movement and Accessibility Needs

There are many affective events that describe our movement or accessibility situations. When we cannot move freely, or access something in a timely manner, then we are impacted negatively. Specifically, this type of needs include: (1) the need to move or change locations freely; (2) the need to access things or services in a timely manner.

Examples

- “*I have been waiting for 5 hours*” is negative because the need to access things in a timely manner is not satisfied.
- “*I was stuck in my car*” is negative because the need to move freely is violated.
- “*The bus is late*” is negative because the need to take public transport on time is violated.

● Emotions, Sentiments, and Opinions

There are also many affective events or state expressions in human language that do not belong to any human need categories, but only directly describe people’s affective states such as sentiments, emotions, or opinions. I define a separate category to group these affective events together, which includes: (1) events that di-

rectly describe experiencers' sentiments, emotions, feelings, or physical expressions of emotions; and (2) events that express an opinion towards an object. I define that if an event both expresses a sentiment/emotion and is also related to a human need category above, I prefer to categorize it into the corresponding human need category. For example, "I love my family" not only expresses a positive emotion, but also indicates a good social relationship. Based on the definition, it is categorized into the *Social Needs* category. As presented in Chapter 1, events studied in this dissertation refer to both dynamic events and stative events. Ideally, state expressions might be separately from dynamic events because most of these sentiment, emotion, and opinion expressions are states. However, recognizing state expressions itself is difficult research problem. For example, it is hard to distinguish stative expressions for expressions such as "he got angered", "he was angered by the answer", "he was worried", and "he feels worried".

Examples

- "I am happy" is positive because it describes a positive internal emotion state.
- "I smiled" is positive because it describes a happy emotion.
- "Canadians are good" is positive because it describes a positive opinion.

● None of the Above

There are some affective events that do not belong to any of previous human need categories or the *Emotions* category. I categorize all of these affective events in the *None of the Above* category, which includes, but is not limited to (1) events or situations that are too general or abstract to be assigned to any previous categories; or (2) the reason why the event is positive or negative falls into a different category than the ones listed above.

Examples

- "I had a problem" is negative, but we do not know the specific reason why.
- "I made a mistake" is negative but we do not know what the mistake is.

- “*I am powerless*” is negative because the need to have power is violated. It is categorized into this class because the human need implied by this event cannot be classified into any previous human need categories.
- “*I lost authority*” is negative because the need to have authority is not satisfied. The type of human need cannot be categorized into any previous human need classes either.

6.1.2 Gold Human Need Annotations and Analysis

I hypothesized that the reason for events being affective can often be explained based on the satisfaction or violation of human needs. However, two questions remain to be answered about this hypothesis. First, what percentage of affective events can be classified into the 7 human need categories? Second, can human annotators consistently agree on assigning a human need category to an affective event?

To answer these two questions, I conducted a study to add human need annotations to the 559 affective events that were manually identified as positive or negative from a random set of events (details about the affective polarity annotations are presented in Section 3.3). I asked three human annotators to assign the most appropriate category label to each affective event based on the definitions of human need categories (the annotation guidelines are presented in Appendix E). I measured their pairwise inter-annotator agreements using Cohen’s kappa, which were $\kappa=.69$, $\kappa=.66$ and $\kappa=.65$. This demonstrates that human annotators can achieve good agreements ($\kappa \geq .65$) on this task. I then assigned the majority category label to each event as the gold category label. I found that 17 affective events were assigned three different labels, so I discarded them because annotators did not agree on these cases. I ended up with a gold standard data set of 542 affective events with human need category labels. Some examples are shown in Table 6.2.

The distribution of human need categories is shown in Table 6.3, which reveals that the majority (58%) of the affective events can be categorized into the 7 human need categories. Twenty-four percent of the affective events are simply sentiment or emotion expressions (e.g., “I’m happy”, “I hate it”). If we remove pure sentiment or emotion expressions, then 76% of the remaining affective events can be categorized into the 7 human need categories. Table 6.3 also shows that 18% of affective events were assigned with a *None of the Above*

Table 6.2: Affective Event Examples with Human Needs Category Labels.

Positive Events	Human Need
⟨ I, take, advantage, of breakfast ⟩	Physiological
⟨ ear, be, better, - ⟩	Health
⟨ I, watch, Hellboy II, - ⟩	Leisure
⟨ we, get, marry, - ⟩	Social
⟨ I, get, my new laptop, - ⟩	Finance
⟨ my memory, be, vivid, - ⟩	Cognition
⟨ my heart, feel, happy, - ⟩	Emotion
⟨ we, be, legal, - ⟩	None
Negative Events	Human Need
⟨ I, grow, hungry, - ⟩	Physiological
⟨ my face, look, pale, - ⟩	Health
⟨ -, rain out, game, - ⟩	Leisure
⟨ girl, laugh, -, at me ⟩	Social
⟨ house phone, not work, -, - ⟩	Finance
⟨ I, lose, attention, - ⟩	Cognition
⟨ I, be, scared, - ⟩	Emotion
⟨ it, not work, -, for me ⟩	None

Table 6.3: Distribution of Human Need Categories (each cell shows the frequency and percentage).

Physiological	Health	Leisure	Social	Finance
19 (4%)	52 (10%)	75 (14%)	108 (20%)	29 (5%)
Cognition	Freedom	Emotion	None	
26 (5%)	7 (1%)	128 (24%)	98 (18%)	

label. I conducted a further analysis on these affective events, and found three reasons for why some events were classified to the *None* category. First, some events describe very abstract positive or negative events where we know the polarity but not the specific situations (e.g., “we have trouble”, “it does not work for me”). Second, some meanings of events are not clear, which could be caused by parsing or extraction errors (e.g., “bit beats up”, “I fell things”). Third, some events are affective because of other nuanced human need categories. For example, “I’m powerless” is negative because people have needs to be strong, to have power and authority. “I’m legal” is positive because people have needs to be in conformity with or permitted by law. Events related to other human needs can be pervasive in the data of other domains (e.g., legal related documents), but these nuanced human needs are rare in the affective events studied in this research. It can be worth

creating new human need categories when studying events in other domains.

Table 6.3 also shows that the *Freedom* category (i.e., Freedom of Movement and Accessibility) only contains 7 instances. I concluded that this class is too small to provide sufficient evaluation data. Therefore, I merged the Freedom category into the None category in the subsequent study and experiments.

Table 6.4 shows the polarity distribution for each human need category. The number in the parentheses following each human need category is the total number of affective events assigned with the category label. Each cell shows the number and percentage of positive (POS) or negative (NEG) events for each category. Table 6.4 shows that Physiological, Health, and None categories have more negative events than positive events. 83% percentage of the affective events in the Health category are negative. The Leisure and Social categories have more positive events. Eighty-seven percent of affective events in Leisure are positive. This demonstrates that the polarity frequencies for each human need category can vary. This unbalanced distribution reflects the human need categories observed in the personal story corpus. For example, in the personal blog posts, people are more likely to discuss positive events related to Leisure rather than negative ones. However, it is not known whether this generally reflects what people blog about, and this question can be worth studying in future work.

Figure 6.1 shows the confusion matrix between the two human annotators (A1 and A2) whose agreement is $\kappa=0.66$. In the figure, the category names are abbreviated as Physiological (Phy), Health (Hlth), Leisure (Leis), Social (Socl), Finance (Fnc), Cognition (Cog), and Emotion (Emo). #Tot denotes the total number of events in each row or column.

Table 6.4: Distribution of Affective Polarities under Different Human Need Categories

Physiological (19)		Health (52)		Leisure (75)	
POS 7 (37%)	NEG 12 (63%)	POS 9 (17%)	NEG 43 (83%)	POS 65 (87%)	NEG 10 (13%)
Social (108)		Finance (29)		Cognition (26)	
POS 73 (68%)	NEG 35 (32%)	POS 14 (48%)	NEG 15 (52%)	POS 12 (46%)	NEG 14 (54%)
Freedom (7)		Emotion (128)		None (98)	
POS 3 (43%)	NEG 4 (57%)	POS 66 (52%)	NEG 62 (48%)	POS 37 (38%)	NEG 61 (62%)

A1 \ A2	Phy	Hlth	Leis	Socl	Fnc	Cog	Emo	None	#Tot
Phy	15	3	0	0	0	0	1	1	20
Hlth	2	36	0	5	0	0	2	3	48
Leis	2	1	60	9	4	1	5	6	88
Socl	0	5	1	85	0	0	1	6	98
Fnc	2	0	0	2	22	0	0	2	28
Cog	0	1	1	4	0	20	3	3	32
Emo	0	3	2	9	2	3	90	12	121
None	2	3	4	13	5	2	22	73	124
#Tot	23	52	68	127	33	26	124	106	559

Figure 6.1: Confusions between Two Annotators on Assigning Human Need Labels.

Each row denotes the number of affective events that were annotated with a label (e.g., “Phy”) by annotator A1. Each column denotes the number of affective events that were annotated with a label (e.g., “Hlth”) by annotator A2. For example, the cell (“Phy”, “Hlth”) denotes that there are 3 affective events that were labeled “Phy” by A1 but “Hlth” by A2. The confusion table shows that human annotators often confused Emotion and None. Further analysis reveals that annotators have difficulty distinguishing events that explicitly express sentiments, emotions from events that are highly associated with sentiments or emotions. For example, “we won” does not explicitly express a sentiment, but it is highly associated with a positive sentiment. The confusion matrix also shows that annotators had disagreements on assigning Social and Leisure labels. This is because some activities are associated with both Social and Leisure needs. For example, annotators may have different understandings about what human needs are satisfied by the event “we have a party”. Some may think that this event is mainly related to social relations because having a party is to be with friends or other people, but others may believe that the main goal of having a party is to have fun. In future, it is worth studying to allow an affective event to be associated with multiple human need categories.

6.2 Categorizing Human Needs with Labeled and Unlabeled Data

Automatically categorizing affective events in text based on human needs is a new task, so I investigated several types of approaches. I began by exploring supervised machine learning classifiers by modeling different types of features. Then, I designed two self-supervised models to exploit a large set of unlabeled events. This section presents the

details of these methods.

First, I created a supervised classifier to categorize affective events using the words in event expressions. I will call this type of classifier an *Event Expression Classifier*. To represent words in event expressions, I explored lexical features, word embedding features, and semantic category features. I also explored several types of machine learning algorithms.

The task is to determine the human need category of an affective event based on the meaning of the event itself (i.e., independent of any specific context), which is the most common interpretation and default meaning of an event in the absence of context. However, I hypothesized that collecting the contexts around instances of events can also provide valuable information for predicting human need categories. Therefore, I also designed an *Event Context Classifier* to exploit the sentence contexts around event instances.

As shown in the previous section, the gold standard data set is relatively small, so supervised learning that relies entirely on manually labeled data may not have sufficient coverage to perform well across the human need categories. However, there are a very large set of events that were extracted from the same blog corpus, but not manually labeled with human need labels. Consequently, I explored two weakly supervised learning methods to exploit the large set of unlabeled events. First, I tried self-training to iteratively improve the event expression classifier. Second, I designed a co-training model that takes advantage of both an event expression classifier and an event context classifier to learn from the unlabeled events. These two types of classifiers provide complementary views of an event, so new instances labeled by one classifier can be used as valuable new data to benefit the other classifier, in an iterative learning cycle.

In the following sections, I will describe each type of approach in detail.

6.2.1 Supervised Classification Models

One straightforward approach for categorizing affective events is to train supervised classifiers using labeled data. I explored the hypothesis that both event expression features and context features can be useful for predicting the human need categories of affective events. Based on this hypothesis, I designed two types of classifiers: *Event Expression Classifiers* that train classifiers using features from event expressions, and *Event Context Classifiers* that build classifiers using features from sentence contexts surrounding event

mentions. The details of creating these two types of classifiers are described in this section.

6.2.1.1 Event Expression Classifiers

To represent event expressions as features, I designed three types of features: *bag-of-words features*, *event embedding features*, and *lexical semantic features*. The bag-of-word features use the words in an event expression as features for recognizing its human need category (e.g., {ear, be, better} for the event <ear, be, better, ->). Each bag-of-word feature is associated with a binary presence value (i.e., 1 or 0). To obtain embedding features of events, I create an embedding vector for an event by computing the average of embedding vectors of words in the event representation. I used the 200-dimensional GloVe (Pennington et al., 2014) word embeddings pretrained on 27B tweets. I also designed semantic features using the lexical categories in the LIWC lexicon (Pennebaker et al., 2007) to capture a more general meaning for each word. LIWC is a dictionary of words associated with “psychologically meaningful” lexical categories, some of which are directly relevant to the task of human need categorization, such as AFFECT, SOCIAL, MONEY, and INSIGHT. I identify the LIWC category of the head word of each component in an event representation and use them as *Semantic Category* features.

To investigate the effectiveness of different classification models, I experimented with three types of classification algorithms: logistic regression (LR), support vector machines (SVM), and recurrent neural network classifiers (RNN). For LR and SVM, I train a one-vs.-rest (ovr) binary classifier for each category, and predict the human need category of an event as the category with the highest confidence score, if the event is classified into more than one category. One advantage of the RNN is that it considers the word order in an event expression, which can be important. In the experiments, I used the Scikit-learn implementation (Pedregosa et al., 2011) for the LR classifier, and LIBSVM (Chang and Lin, 2011) with a linear kernel for the SVM classifier. For the RNN, I used the example LSTM implementation from Keras github (Chollet et al., 2015), which was developed for sentiment analysis. LSTM networks (Hochreiter and Schmidhuber, 1997) are a specific implementation of RNN, and were designed to solve the long-term dependency problem by allowing for information flowing to next state unchanged. In the experiments, I used the default parameters in these implementations.

6.2.1.2 Event Context Classifiers

The event representations in this research were originally extracted from a large collection of blog posts, which contain many instances of the events in different sentences. I hypothesized that the contexts surrounding instances of an event can provide strong clues for identifying the human need category associated with the event. Therefore, I also created *Event Context Classifiers* to exploit the sentence contexts around event mentions. I designed the following 4 event context classifiers by using different types of sentence contexts and methods for extracting features from the sentence contexts:

Context^{SentBOW} : To build this classifier, I use each sentence mentioning an event as a training instance. Specifically, for each event in the training set, I first collect all sentences mentioning this event and assign each of these sentences with the event's human need category label. For each sentence, I extract bag-of-word features (i.e., unigrams) from the sentence, and each bag-of-word feature is assigned a binary presence value. Then, each sentence and its label is used to train this type of event context classifier.

Context^{SentEmbed} : This variation also uses a sentence as a training instance and labels sentences exactly the same way as the previous model, but rather than using bag-of-word features, this method represents each sentence as a dense embedding vector, which is computed as the average of the embeddings for each word in the sentence. In the experiments, I used the 200-dimensional GloVe (Pennington et al., 2014) word embeddings pretrained on 27B tweets.

Context^{AllBOW} : In this classifier, instead of treating each sentence as a training instance, I merge all of the sentence contexts of an event to be a single training instance. Specifically, for an event, I first collect all of the sentences that mention this event, and aggregate them into one giant context for the event. Then, I assign the human need category label of the event as the label for this giant context and use them together as one training instance for this classifier. To train this classifier, I represent each giant context using bag-of-word features.

Context^{AllEmbed} : This variation aggregates the sentences mentioning an event exactly like the previous model, but in this model, I use embedding features instead of bag-of-word features. To compute the embedding of an aggregate context for an event, I first compute an embedding vector for each sentence as the average of the embeddings of its

words. Then, I compute a single context embedding for the giant context by averaging all of the sentence embeddings.

In the personal story blog posts, I noticed that some events appear in many sentences, while others appear in just a few sentences. Since I label each event context instance with the event’s label, the varying size of sentence contexts for an event can break the category distribution in the original training data and lead to worse performance. To maintain a balanced training set of context instances for each event, I randomly sample 10 sentences for each event to use as its contexts. When there are less than 10 sentence contexts for an event, I use all of them.

To predict the human need category of an event, I apply event context classifiers to contexts that mention the event, and compute the human need category label using the prediction results. With different types of event context classifiers, the human need category label for an event is computed differently. Specifically, for the $\text{Context}^{\text{AllBOW}}$ and $\text{Context}^{\text{AllEmbed}}$ models, each event is only associated with a single giant context instance, so the models can directly predict the human need category label of an event by using the giant context associated with the event as input. For the $\text{Context}^{\text{SentBOW}}$ and $\text{Context}^{\text{SentEmbed}}$, each event is associated with multiple sentence contexts. I first apply these two classifiers on the sentence contexts of an event and obtain a single probability distribution for each event by computing the average of probability distributions of these sentence contexts. Finally, I assign each test event with the human need category that has the highest mean probability.

6.2.2 Semi-Supervised Models

Since the gold standard data set is relatively small, supervised classifiers that rely on manually labeled data may not have sufficient coverage to perform well. However, there is a large set of unlabeled events extracted from the same blog corpus as the gold data. Therefore, in this section, I present two types of semi-supervised models, self-training, and co-training, to exploit the effectiveness of using a large set of unlabeled events.

6.2.2.1 Self-Training the Event Expression Classifier

Obtaining more manual annotations is expensive and time consuming. However, it is easy to automatically collect labels for unlabeled events using a previously trained model. These automatically predicted labels can be useful to improve the performance

of the original model. Based on this idea, I designed a self-training model that tries to iteratively improve the event expression classifier by exploiting the unlabeled event data. In this part of the research, I only self-trained the event expression classifier because I observed the event expression classifier performed better than the event context classifier in the experiments (evaluation details can be found in Section 6.3).

Specifically, the self-training process works as follows. First, an event expression classifier is trained using the manually labeled events. Then, the classifier is applied to the unlabeled events and assigns a human need category to each event with a confidence value. For each human need category, I select the unlabeled event that has been assigned to that category with the highest confidence. Therefore, each category will have one additional labeled event at each iteration. The newly labeled events are added to the labeled data set, and the classifier is re-trained for the next iteration. After a number of training iterations, the final self-training classifier is expected to perform better than the initial event expression classifier.

6.2.2.2 Co-Training with Event Expression and Event Context Classifiers

The event expression classifiers use the event expressions to assign human need category to an event, and the event context classifiers use information from the contexts mentioning event instances. These two types of classifiers have different views of an event. I hypothesize that these two views are complementary to each other, and the unlabeled events that are assigned predicted labels and selected confidently by one classifier can benefit another classifier. Therefore, I designed a co-training model to exploit these complementary types of classifiers to iteratively learn from unlabeled data.

Figure 6.2 shows the architecture of the co-training model. To train the co-training model, an event expression classifier and an event context classifier are independently trained on the manually labeled training data. Each classifier is then applied to the large collection of unlabeled events E_U . For each human need category, I then select the event that has been assigned to the category with the highest confidence value as a new instance to label. Consequently, each category will receive two additional labeled events at each iteration, one from the event expression classifier and another one from the event context classifier. In my implementation, the event expression classifier first selects from unlabeled

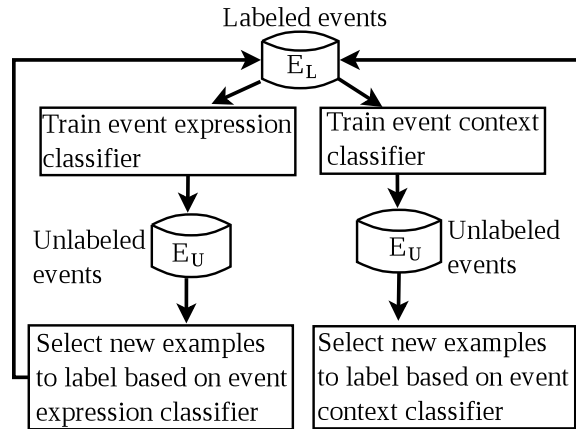


Figure 6.2: The Co-Training Model for Human Needs Categorization

events, then the event context classifier does the selection. If an event is selected previously, it will not be selected again. This ensures that there are 16 new events in total at each iteration.

Both sets of newly labeled events are then added to the labeled set E_L , and each of the classifiers is re-trained on the expanded set of labeled data. Because the classifiers have different views of the events, the new instances labeled by one classifier serve as fresh training instances for the other, unlike self-training with a single classifier where it is learning entirely from its own predictions.

The detailed co-training algorithm is shown in Algorithm 3. The input to the algorithm are the sets of labeled events E_L and unlabeled events E_U . Each event is associated with both an event expression and the set of sentences in which it occurs in the blogs corpus.

For each iteration, the event expression classifier is first trained using the labeled events E_L with the event expression view. Then, I construct an event context view X_{con} for each event in the labeled set E_L . The context sentences are used differently depending on the type of context model (described in Section 6.2.1.2). An event context classifier is then trained using the context view X_{con} . Both classifiers are then independently applied to the unlabeled events E_U . For each human need category, each classifier selects one event to label based on its most confident prediction. All of the newly labeled events are then added to the labeled training set E_L , and the process repeats.

The co-training process simultaneously trains two classifiers, so here I explain how the resulting classifiers are used for predicting human need categories after the co-training

Algorithm 3 Co-Training Algorithm

- 1: **Input:** Labeled E_L , unlabeled E_U events
 - 2: **while** Not maximum iteration **do**
 - 3: Train the event expression classifier on E_L
 - 4: Construct context view (X_{con}) of E_L
 - 5: Train the event context classifier on X_{con}
 - 6: Apply the event expression classifier to E_U and select new labeled events (E_{exp})
 - 7: Apply the event context classifier to E_U and select new labeled events (E_{con})
 - 8: Update labeled events:

$$E_L = E_L \cup E_{exp} \cup E_{con}$$
 - 9: **end while**
-

process has finished. For each event e in the test set, I apply both the event expression classifier and the event context classifier, which are logistic regression models that produce a probability distribution over the human need categories. Then, I explore two different methods to combine the two probability distributions for each test event: (1) **sum**, I compute the final probability vector $p(e)$ by applying the element-wise summarization operation to the two predicted probability vectors; (2) **product**, I compute the final $p(e)$ as the element-wise product of the two vectors. Then, the final probability vector is normalized to make sure the sum of probabilities over all classes is 1. Finally, I predict an event’s human need category as the one with the highest probability.

6.3 Evaluation

In this section, I present experiments to evaluate the methods described in the previous section. For all of the experiments, I use the gold standard data described in Section 6.1. Experimental results are reported based on 3-fold cross-validation on the 542 affective events that are manually labeled with human need categories. In the experiments, I used classification models with default parameters, and did not tune any of the models. In the following, I first present the evaluation metrics used in the experiments. Then, I evaluate and show the performance of a baseline that uses the LIWC lexicon, the event expression classifiers, the event context classifiers, and two semi-supervised models: self-training and co-training. Finally, I present an analysis by showing a confusion matrix between

gold labels and predicted labels by the best performing system, and examples that were incorrectly classified by the best system.

6.3.1 Evaluation Metrics

In the following experiments, I use two evaluation methods to measure the performance of each human need categorization method. The first evaluation method is called **FoldAvg**, which computes the final performance score for each method as the average of the scores on the 3 test sets from the 3 folds. For each fold, I first compute the precision, recall, and F1 score for each human need category, then obtain the macro-average precision, recall, and F1 on the test set (i.e., averaging across the human need categories). Since I use 3-fold cross-validation, there are 3 sets of macro-averaged scores. I compute the final precision, recall, and F1 for each system as the average across the 3 folds. Since FoldAvg reports the average precision, recall, and F1 score over 3 folds, the F1 score can be smaller than both precision and recall in some cases. In the paper (Ding and Riloff, 2018a), the FoldAvg evaluation metric was used to report the performance of each system.

The second evaluation method is called **TestAvg**, which directly computes the performance score for each system on the test instances across 3 folds. For each system, I first obtain the predictions on the test set in each fold, and then merge all of the predictions into one single evaluation set. Then, I compute the precision, recall, and F1 score for each category on the test instances. Finally, I compute a system's performance as the macro-average precision, recall, and F1 across all of the categories. For the sake of comparison, I present results for both evaluation metrics below.

6.3.2 LIWC Lexicon Baseline

The LIWC lexicon (Pennebaker et al., 2007) is a dictionary of words associated with various lexical categories. Besides pronominal and emotion categories, it also contains cognition and psychology categories of words, which are closely related to the human needs categorization task. Consequently, the LIWC lexicon can be used for this task as a valuable baseline to access the performance of existing resources. I first analyzed the categories in LIWC, and manually built a mapping from LIWC categories to the human need categories, which is shown in Table 6.5. To predict the human need category of an event, I designed a rule-based system that first looks up the LIWC category of each word in

Table 6.5: LIWC Mapping to Human Need Categories.

LIWC Category		Human Need Category
Ingest	→	Physiological
Health, Body, Death	→	Health
Leisure	→	Leisure
Social	→	Social
Money, Work	→	Finance
Inhib, Insight	→	Cognition
Affect	→	Emotion

an event, maps it to a human need category, and then uses the majority category across all words in the event expression as the final human need category. For the arguments of an event, in most cases, the Agent is a pronoun and does not contribute much to the meaning of the event. In many cases, the Prepositional Phrase (PP) is also optional. Therefore, if there is a tie, I remove a component one by one in the order of Agent, PP, Theme until it can produce a majority label. I do not remove the Predicate because the Predicate is an important component to an event in most cases. If none of the words in an event are contained in LIWC, or their categories cannot be mapped to our categories, then I assign a *None of the Above* category label to the event.

6.3.3 Performance of the Event Expression Classifiers

Table 6.6 shows the results for the LIWC lexicon-based method and the event expression classifiers. The top row of Table 6.6 shows that LIWC achieved 39% with 47.7% precision based on the FoldAvg metric. The reason is that some categories in LIWC are more general compared with the definitions of the corresponding human need categories. For example, the words “abandon” and “damage” belong to the Affect category (corresponding to the Emotion category) in LIWC. However, based on the human need definitions, the event “my house was damaged” actually belongs to the Finance category.

The LR and SVM rows in Table 6.6 show the performance of the logistic regression (LR) and support vector machine (SVM) classifiers, respectively. I evaluated classifiers with bag-of-words features (BOW) and classifiers with event embedding features (Embed), computed as the average of the embeddings for all words in the event expression. I also tried adding semantic category features from LIWC to each feature set, denoted as +SemCat. The FoldAvg and TestAvg results show that the Embed features performed best

Table 6.6: Performance of LIWC Baseline and Event Expression Classifiers

Method	FoldAvg			TestAvg		
	Precision	Recall	F1	Precision	Recall	F1
LIWC	47.7	39.0	38.6	46.6	38.9	42.4
LR ^{BOW}	33.6	28.7	27.3	37.8	28.6	32.6
LR ^{BOW+SemCat}	55.2	39.6	41.9	52.6	39.6	45.2
LR ^{Embed+SemCat}	60.1	49.3	51.9	58.3	49.4	53.5
LR ^{Embed}	64.2	51.7	54.8	62.4	51.8	56.6
SVM ^{BOW}	52.3	43.1	44.8	51.4	43.0	46.8
SVM ^{BOW+SemCat}	51.0	45.9	46.8	49.4	45.8	47.6
SVM ^{Embed+SemCat}	50.4	48.4	48.6	49.9	48.6	49.2
SVM ^{Embed}	51.3	50.7	50.5	51.2	50.8	51.0
RNN ^{Words}	45.2	39.6	40.1	41.3	39.6	40.4
RNN ^{EmbedSeq}	58.0	53.7	54.4	57.2	53.9	55.5

for both the LR and SVM classifiers. Adding the SemCat features improved upon the bag-of-words representations, but not the embeddings.

The last two rows of Table 6.6 show the performance of two RNN classifiers, one using lexical words as input (RNN^{Words}) and one using pretrained word embeddings as input (RNN^{EmbedSeq}). The RNN^{EmbedSeq} system takes the sequence of word embeddings as input rather than the average embeddings. As with the other classifiers, the pretrained word embedding features performed best, achieving FoldAvg F1 score 54.4% and TestAvg F1 55.5%. RNN^{EmbedSeq} achieved comparable F1 score to that of the LR^{Embed} system when using the FoldAvg evaluation metric. However, the TestAvg F1 score of RNN^{EmbedSeq} is 1.1% lower than that of the logistic regression model, and the precision of the LR^{Embed} is much higher (6.2% higher based on FoldAvg and 5.2% higher based on TestAvg) than the neural net model. This demonstrates that the logistic regression models with pretrained word embedding features perform better than the neural net models. Because neural net models often seem to benefit from large training sets, the relatively small size of the training data may not be ideal for an RNN.

Overall, I concluded that the logistic regression classifier with event embedding features (LR^{Embed}) achieved the best performance because of its F1 score and higher precision.

6.3.4 Performance of the Event Context Classifiers

Table 6.7 shows the performance of the event context classifiers described in Section 6.2.1.2. Since logistic regression worked best in the previous experiments, I only evaluated logistic regression classifiers in the remaining experiments. The results show that using each context sentence as an individual training instance ($\text{Context}^{\text{SentBOW}}$ and $\text{Context}^{\text{SentEmbed}}$) substantially outperformed the classifiers that merged all the context sentences as a single training instance ($\text{Context}^{\text{AllBOW}}$ and $\text{Context}^{\text{AllEmbed}}$). Overall, the best performing system $\text{Context}^{\text{SentEmbed}}$ achieved an F1 score of 44.3% with 59.1% Precision using the FoldAvg metric. In addition, I also created a classifier that combined event expression features and event context features together, but the classifier did not improve performance.

Since the event context classifiers achieved (roughly) similar levels of precision to the event expression classifiers, and these two types of classifiers represent complementary views of events, a co-training framework seemed like a logical way to use them together to gain additional benefits from unlabeled event data.

6.3.5 Performance of Self-Training and Co-Training Models

In this section, I evaluate the semi-supervised self-training and co-training methods that additionally use unlabeled data. To keep the number of unlabeled events manageable, I only used the unlabeled events that had frequency ≥ 100 , which produced an unlabeled data set of 23,866 events. All these unlabeled events were extracted from the story blog corpus, which is described in Section 3.1. Same as the events with gold labels, unlabeled events are extracted using the enhanced event frame representation whose extraction method is presented in Section 3.2.

I used the best performing LR^{Embed} model as the event expression classifier in these models, and the co-training framework includes the best performing event context clas-

Table 6.7: Performance of Event Context Classifiers

Method	FoldAvg			TestAvg		
	Precision	Recall	F1	Precision	Recall	F1
$\text{Context}^{\text{AllBOW}}$	20.6	18.0	17.8	22.5	18.0	20.0
$\text{Context}^{\text{AllEmbed}}$	38.2	29.9	29.1	38.0	29.9	33.5
$\text{Context}^{\text{SentBOW}}$	48.2	31.4	32.8	63.2	31.3	41.8
$\text{Context}^{\text{SentEmbed}}$	59.1	41.9	44.3	66.6	42.0	51.5

sifier ($\text{Context}^{\text{SentEmbed}}$) as well. I also experimented with the **sum** and **product** variants for co-training (described in Section 6.2.2.2), which are denoted as $\text{CoTrain}^{\text{sum}}$ and $\text{CoTrain}^{\text{prod}}$. In the experiments, I ran both the self-training and co-training methods for 20 iterations. I also conducted experiments to run these two methods for more iterations, but did not observe much improvement.

Figure 6.3 and Figure 6.4 track the performance of the self-training and co-training models after each iteration, in terms of FoldAvg F1 score and TestAvg F1 score. The flat lines in both figures show the performance for the best classifier that uses only labeled data (LR^{Embed}). Both types of semi-supervised models yield performance gains from iteratively learning with the unlabeled data, but the co-training models perform substantially better than the self-training model. Even after just 5 iterations, co-training achieves an F1 score over 59% using FoldAvg and F1 over 61% using TestAvg, and by 20 iterations, performance improves to 61% FoldAvg F1 and 62.1% TestAvg F1 score.

Table 6.8 shows the results for these models after 20 iterations, which was an arbitrary stopping criterion, and after 17 iterations, which happened to produce the best results for all three systems using both FoldAvg and TestAvg metrics (please note that the self-training model achieved its best score of 58% at 12, 17, and 20 iterations when using the TestAvg metric). The first two rows show the results of the best performing event context

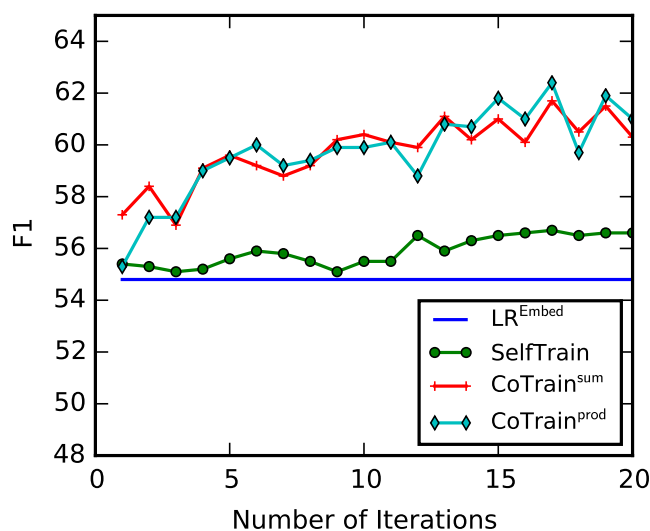


Figure 6.3: Learning Curves of Self-Training and Co-Training Using the FoldAvg Metric

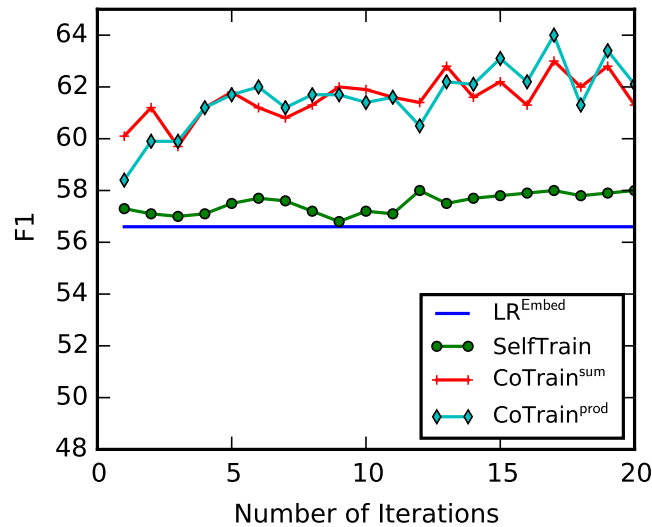


Figure 6.4: Learning Curves of Self-Training and Co-Training Using the TestAvg Metric

classifier ($\text{Context}^{\text{SentEmbed}}$) and best performing event expression classifier (LR^{Embed}) from the previous experiments, for the sake of comparison.

Table 6.8 shows that after 20 iterations, the $\text{CoTrain}^{\text{prod}}$ model performed best. Furthermore, the co-training model improves upon both the precision and recall.

All three systems performed best after 17 iterations, so I show those results as well to give an idea of additional gains that would be possible if we could find an optimal stopping criterion. The gold annotation data set was small so I did not feel that I had enough

Table 6.8: Performance of Self-Training and Co-Training

Method	FoldAvg			TestAvg		
	Precision	Recall	F1	Precision	Recall	F1
<i>Supervised Models</i>						
$\text{Context}^{\text{SentEmbed}}$	59.1	41.9	44.3	66.6	42.0	51.5
LR^{Embed}	64.2	51.7	54.8	62.4	51.8	56.6
<i>After 20 Iterations</i>						
SelfTrain	63.2	54.2	56.6	62.0	54.4	58.0
$\text{CoTrain}^{\text{sum}}$	66.2	58.2	60.3	64.7	58.3	61.3
$\text{CoTrain}^{\text{prod}}$	67.1	58.7	61.0	65.7	58.9	62.1
<i>Best Results, After 17 Iterations</i>						
SelfTrain	63.5	54.1	56.7	62.3	54.3	58.0
$\text{CoTrain}^{\text{sum}}$	68.6	59.0	61.7	67.4	59.2	63.0
$\text{CoTrain}^{\text{prod}}$	69.7	59.5	62.4	69.0	59.7	64.0

data to fine-tune parameters, but there is a potential to further improve performance given additional tuning data.

Table 6.9 shows a breakdown of the performance across the individual human need categories for two models: the best event expression classifier and the best co-training model (CoTrain^{Prod} after 17 iterations). The results in this table are reported using the FoldAvg metric. The table shows that the co-training model outperformed the LR^{Embed} model on every category. Co-training improved performance the most for the Finance and Cognition categories, yielding F1 score gains of +12% and +16%.

6.3.6 Analysis

I manually examined the predictions of the best system (i.e., CoTrain^{Prob}) to better understand its behavior. I found that most of the correctly classified Physiological events were related to food, while the correctly classified Cognition events were primarily about learning and understanding. The co-training model missed many events for the Health, Finance, and Cognition classes. For Health, many medical symptoms were not recognized, such as *“my face looks pale”* and *“I puked”*. For Finance, the system missed events related to possessions (e.g., *“engine stopped running”* and *“my clock is wrong”*) and jobs (e.g., *“I went to resign”*).

I also took a closer look at which categories were confused with other categories. Figure 6.5 shows the confusion matrix between CoTrain^{Prod} and the gold annotations. Each cell shows the total number of confusions across the test sets of 3 folds. In the figure, Predict

Table 6.9: Breakdown of Results across Human Need Categories. Each Cell Shows Precision, Recall, and F1.

Category	FoldAvg					
	LR ^{Embed}			CoTrain ^{Prod}		
	Pre	Rec	F1	Pre	Rec	F1
Physiological	82	57	67	81	68	74
Health	65	40	49	68	50	57
Leisure	62	59	60	69	63	66
Social	61	72	66	68	79	73
Finance	61	31	40	67	44	52
Cognition	75	31	42	92	46	58
Emotion	60	75	66	64	74	69
None	47	49	48	48	52	50

Predict \ Gold	Phy	Hlth	Leis	Socl	Fnc	Cog	Emo	None	#Tot
Phy	13	1	0	0	1	0	0	2	17
Hlth	1	26	1	0	1	1	4	8	42
Leis	1	1	48	4	0	1	4	10	69
Socl	0	6	4	84	2	3	10	11	120
Fnc	1	0	2	0	13	0	1	5	22
Cog	0	0	0	0	0	12	1	2	15
Emo	1	5	12	12	3	1	91	16	141
None	2	13	8	8	9	8	17	51	116
#Tot	19	52	75	108	29	26	128	105	542

Figure 6.5: Confusions between Predictions and Gold Human Need Annotations.

denotes predictions and Gold denotes the gold annotations. The category names are abbreviated as Physiological (Phy), Health (Hlth), Leisure (Leis), Social (Socl), Finance (Fnc), Cognition (Cog), and Emotion (Emo). #Tot denotes the total number of events in each row or column.

The co-training model had difficulty distinguishing the None category from other classes, presumably because None does not have its own semantics but is used for affective events that do not belong to any of the other categories. The system also often confuses Emotion with Leisure and Social. This happens because many event expressions contain words that refer to emotions. The annotation guidelines instructed annotators to focus on the event and assign the Emotion label only when no event is described beyond an emotion (e.g., “I was thrilled”). Consequently, the gold label of “I love travel” is Leisure and “I’m worried about my mom” is Social, but both were classified by the system as Emotion. In future work, it may be advantageous to allow event expressions to be labeled as both an explicit Emotion and a Human Need category based on the target of the emotion.

6.4 Chapter Summary

This chapter aims to understand the reason for events being affective and presents a task of categorizing affective events into one of the human need categories. This chapter first demonstrates through a manual annotation study that most of the affective events in blogs can be categorized into a small set of human need categories. Then, this chapter describes a variety of methods for this task including supervised classifiers and semi-supervised models. This chapter also describes a co-training model in detail. Finally, this chapter demonstrates their effectiveness with a thorough experimental analysis. The main

content of this chapter is summarized below.

- This chapter presents a new task of categorizing affective events into human need categories. The human need associated with an affective event can often be used to explain why the event is positive or negative. This chapter developed 7 human need categories by selecting and separating some categories from the two psychology theories (Maslow et al., 1970; Max-Neef et al., 1991), which are: *PHYSIOLOGICAL*, *HEALTH*, *LEISURE*, *SOCIAL*, *FINANCIAL*, *COGNITION*, and *FREEDOM*. This chapter also describes a manual annotation effort to obtain gold human need annotations for a random set of affective events, which are then used for evaluating automatic classification methods. The manual analysis demonstrates that 58% of affective events extracted from blog posts can be categorized into the 7 human need categories.
- This chapter describes a variety of classification models to access the difficulty of the new task of categorizing affective events based on their implied human needs. These methods include supervised event expression classifiers that use features from event expressions, supervised event context classifiers that are trained on event context features, and a self-training model that exploits unlabeled data.
- This chapter presents the details of a co-training model, which improves human need classification performance by taking advantage of the two views provided by an event expression classifier and an event context classifier, and a large set of unlabeled events. The co-training model first trains an event expression classifier and an event context classifier independently on the manually labeled training data, and applies the two classifiers to unlabeled events. Then, highly confident predictions by each classifier are selected and added into the training set. The expanded training set is used to re-train both the event expression classifier and the event context classifier at the next iteration. Finally, the resulting two classifiers are used together to predict the human need category of an affective event.
- This chapter also presents thorough experiments for evaluating the performance of methods designed to categorize affective events. Experimental results show that a logistic regression classifier with event embedding features outperforms a baseline

using LIWC lexicon and other supervised classification models. Results also show that self-training the event expression classifier can slightly improve performance. The co-training model that uses both event expression view and event context view significantly improves the classification performance, which demonstrates that the two views provided by event expression classifiers and event context classifiers are complementary to each other. In addition, this chapter analyzes the behavior of the best performing co-training method by showing that it often confuses Emotion with Leisure and Social, which indicates that it may be advantageous to allow event expressions to be labeled with multiple labels in future work. Further analysis also shows that performance on Health, Finance, and Cognition categories still has substantial room for improvement.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

This dissertation presents research on learning two types of knowledge about affective events: (1) affective polarity of events, and (2) human need categories associated with affective events. Since events are frequently mentioned and discussed in various texts such as news, narrative stories, conversations, and Web blog posts, acquiring these two types of knowledge can help achieve better understanding of how an event impacts people and the reasons why people will feel good or bad after experiencing events. This chapter summarizes the research presented in this dissertation and the contributions, and discusses future research directions based on this research.

7.1 Research Summary and Contributions

In this section, I summarize the research conducted in this dissertation and present the contributions. In Chapter 1, I presented two research claims for this dissertation. In this section, I will revisit the research claims and demonstrate that these claims are supported by the empirical results.

This dissertation consists of two lines of research on learning knowledge of affective events. The first part of the research aims to identify affective events and recognize their affective polarities. The second part tries to recognize the human need category associated with each affective event. The research claim for the first part of work is discussed below.

Claim #1. Many affective events in personal stories can be identified and assigned prior polarities using graph-based semi-supervised learning.

Though there have been many resources and systems that were developed to recognize the sentiment of a given text, in my research, I found that existing sentiment analysis resources are not sufficient to identify affective events and recognize their polarities. In this part of research, I designed two graph-based semi-supervised models to identify affective events and assign prior polarities. The prior polarity of an event is the most typical

understanding of how the event impacts people, which is independent of context.

In Chapter 4, I presented a semi-supervised Event Context Graph (ECG) model that was designed to extract affective events using discourse and event collocation information. Instead of using supervision from manual annotations, the ECG model obtains weak supervision by automatically identifying sentence contexts with strong polarity values with an existing sentiment analysis classifier. Then, the ECG model uses a label propagation algorithm to iteratively spread polarity evidence from affective sentence nodes to event nodes through three types of edges: local context edges, discourse proximity context edges, and event co-occurrence edges. Experimental results presented in Section 4.3 show that the ECG model can obtain better performance than other systems at recognizing affective events, and achieved 90% and 84.5% accuracy on the top100 and top400 affective events, respectively.

In Chapter 5, I described a semi-supervised Semantic Consistency Graph (SCG) model to identify affective events using semantic relations between event expressions. The SCG model obtains weak supervision by assigning each event with an initial polarity value using sentiment resources. Then, the model builds a Semantic Relations Graph with event-event similarity, event-event opposition, and event-component edges, and uses an iterative learning algorithm to infer correct polarities of events by optimizing the semantic consistency in the graph. Experimental results presented in Section 5.3 show that all of the three semantic relations contribute to improving the performance for recognizing affective events. The SCG model outperformed previous methods and achieved a 71.4% F1 score on a random set of events. Further analysis also shows that the SCG model can acquire over 110,000 affective events with >90% precision for positive events and >80% precision for negative events.

Overall, this part of the research demonstrates that graph-based semi-supervised learning models can identify affective events and assign polarities with high precision, achieving better performance than previous resources and systems. In addition, this part of the research produced a large set of affective events with prior polarities that are automatically assigned with high precision. The learned affective events with automatically predicted polarities were released and are freely available for the research community.

Claim #2. Affective events can be automatically categorized into a small set of human

need categories by co-training models with views based on event expressions and event contexts.

When we discuss events in our daily lives, we not only understand how an event affects us but also know why the event impacts us in that way. In this part of my research, I hypothesized that the reason why an event is positive or negative arises from the satisfaction or violation of a human need. In Chapter 6, I defined a new task of categorizing affective events based on their implied human needs and hypothesized that the majority of affective events can be categorized into a small set of human need categories. In a manual annotation study, I defined 7 human need categories derived from the previous research in psychology, and asked human annotators to label a random set of affective events with these categories. The annotation results demonstrate that 58% of all affective events (and 76% of affective events that are not pure sentiment or emotion expressions) can be categorized into these 7 categories, and human annotators achieved good annotation agreement on this task.

In Chapter 6, I presented a co-training model for recognizing human needs of affective events by taking advantage of two views based event expressions and event contexts. The co-training model first trains an event expression classifier and an event context classifier independently on the manually labeled training events. Then, the two classifiers are applied to unlabeled events and highly confident predictions by each classifier are selected and added into the training set. Finally, the two classifiers are re-trained on the expanded training set at the next iteration. Experimental results presented in Section 6.3 show that the co-training model outperformed the best event expression classifier and event context classifier, achieving +7% higher F1 score than the best individual model. Overall, this part of my research demonstrates that co-training models with views based on event expressions and event contexts can predict human need categories of affective events more effectively than the individual classification models.

7.2 Future Directions

In this dissertation, I designed graph-based semi-supervised models to learn affective polarities of events, and a co-training model to recognize human needs associated with affective events. Though these methods achieved much better performance than previous

methods, there is still room for further improvement on both of these tasks. As I am wrapping up this dissertation, I am also thinking several future research directions that are worth exploring, which include jointly learning affective polarity and human needs, studying affective events in narrative stories and conversations, and building a hierarchical knowledge base of affective events. In the following sections, I will discuss each of these potential research directions and the problems that need to be solved in future work.

7.2.1 Jointly Learning Affective Polarity and Human Needs

Experimental results presented in Sections 5.3 and 6.3 show that the SCG model and the co-training model can achieve better performance than previous methods on recognizing affective polarity and human needs of affective events. However, the results also indicate that these methods are not perfect, for example, the SCG model achieved 71.4% F1 score on recognizing the affective polarity of all events, and the co-training model obtained 64% TestAvg F1 score on the task of categorizing affective events into human need categories. These results indicate that more research needs to be explored to improve performance on both of these tasks.

One potential research direction is to exploit the relationship between the tasks of recognizing affective polarity of events and categorizing affective into human need categories. As shown in Table 6.4 in Chapter 6, the positive and negative polarities are not evenly distributed for many categories such as Physiological, Health, Leisure, and Social. For example, 83% of affective events associated with Health needs are negative, and 87% of affective events with Leisure needs are positive. This observation reveals that we are more likely to know the polarity of an event if we know which human need category it is related to. Therefore, understanding the human need category for an event can potentially help predict its polarity. In addition, I also found that the human need category distributions under different polarities are different. For example, as shown in Table 6.3, 10% of affective events are annotated with Health category label. However, the percentage of the health-related affective events increases to 17% if we know the event is negative, and decreases to 3% when it is positive. Similarly, 14% of affective events are associated with the Leisure category, but the percentage increases to 23% when we know it is a positive event, and decreases to 4% when it is negative. This observation shows

that human need categories under different polarities have different distributions, and knowing an event's polarity can potentially help predict the human need category of an event.

As discussed above, the two tasks of recognizing affective polarity and learning the human needs of affective events are related. To jointly learn these two tasks, there are 2 questions remaining to be solved. Since the human need categories are proposed to explain the affective polarity of events, only the affective events (positive or negative) were annotated for this research, not the neutral ones. Therefore, the first question is should we assign human need categories to neutral events? Answering this question requires us to adapt the current definitions of human need categories to neutral events, which at the same time may lead to a deeper understanding of the relationship between affective polarity and human needs associated with events. The second question is how the two tasks can be jointly learned together. The design of a new joint learning model would be a promising direction for future research.

7.2.2 Recognizing Affective Polarity and Human Needs of Events in Stories and Conversations

This dissertation focuses on acquiring affective knowledge of events, independent of context. The ultimate goal is to obtain better understanding of natural language, especially narrative stories and conversations using the learned knowledge. As a step forward, another future research direction is to apply the learned knowledge to analyze the events in narrative stories and conversations. Exploring this research direction can further validate the quality and quantity of the learned prior knowledge of events, i.e., we can evaluate accuracy of the learned knowledge when it is applied to narrative stories and conversations, and we will also assess the coverage of the learned knowledge on identifying affective events in context.

Knowing how a story character is impacted by an event, and the reason, can help us understand why the character has a specific plan or goal. However, several practical questions need to be answered in this potential future research. First, we need to collect a large set of narrative stories and conversations discussing daily life events. I used the Web blog posts in this dissertation, but some of the blog posts are very noisy and difficult to use for this potential research even though I have filtered many non-story blog posts. Second,

manual annotation for this research can also be hard for human annotators because of the difficulty of the research problem itself. Annotation guidelines need to be created carefully and annotators need to be trained to achieve good annotation agreement.

More importantly, the challenge for this research problem is to design effective methods to understand the *contextual polarity* and *contextual human needs* associated with event mentions. The contextual polarity and contextual human needs associated with an event are determined by both the meaning of the event and its surrounding context, which is different from the prior polarity and human needs studied in this dissertation, whose values are based on the meaning of the event and world knowledge. There has been prior research focusing on understanding the contextual polarity of phrases (Wilson et al., 2005). Understanding the contextual polarity can be difficult because it can be influenced by its context. For example, in the description “We went to Disney World last weekend, but we had a very bad time there.”, the contextual polarity of the event “went to Disney World” is negative even though it is a typically enjoyable event. Similarly, the contextual human needs associated with event instances can also be influenced by its context. For example, in the description “I lost my girlfriend’s ipad. She was angry and wanted to break up with me.”, if we do not consider the context, the event “lost my girlfriend’s ipad” is related to the Finance needs based on world knowledge. However, the contextual human needs associated with the event are more about the Social needs rather than just Finance when considering its context, because the event eventually caused a bad social relation, which is more important for the experiencer.

7.2.3 Building a Hierarchical Knowledge Base of Affective Events

In this research, events are represented using frame-like structures that consist of 4 components: Agent, Predicate, Theme, and Prepositional Phrase. To obtain the most generalized event representations, all events are normalized with two processing steps: (1) events expressed in passive voice sentences are normalized to active voice, (2) all words in event representations are lemmatized. These two normalization steps can normalize events in different voice constructions, tenses, and word forms. However, many events with very similar meanings are still represented as different events. For example, the meanings of <I, eat, cake, ->, <I, eat, chocolate cake, ->, and <I, eat, chocolate, -> are

similar but they are treated as totally different events in the current knowledge base (i.e., the AffectEventKB described in Chapter 5). One potential future research direction is to build a hierarchical knowledge base of affective events, in which the bottom level events are specific and higher level events will be more generalized. For example, the event $\langle I, \text{eat, chocolate cake, -} \rangle$ can be generalized to $\langle I, \text{eat, \#SweetFood\#, -} \rangle$ in which “#SweetFood#” is a more generalized concept than “chocolate cake”. Then, this generalized event can be matched with many other similar events. The main advantage of this hierarchical knowledge base of affective events is that it could back off to generalized events when an event cannot find any match among the lower level specific events in the knowledge base. For example, if the event $\langle I, \text{eat, chocolate pie, -} \rangle$ cannot find any match in the low level specific events, then we can generalize the event to $\langle I, \text{eat, \#SweetFood\#, -} \rangle$ and find a match in the knowledge base.

The challenge for this potential future research is to obtain generalized representations of events and preserve their original polarities at the same time (i.e., the generalized events should have same polarity as the specific events). The first question is how to build a hierarchy of concepts. Manually building a concept hierarchy can be expensive and inflexible to expand in the future. Though some existing knowledge bases (e.g., WordNet) may have concept hierarchies, many of these concepts are too abstract to preserve the polarity of events. One potential key step to solve this problem is to design an effective automatic method to learn and build the concept hierarchy from the data. The second question is how to accurately generalize each event component to a generalized concept in the concept hierarchy. In the enhanced event frame representation described in Section 3.2, the Agent, Theme, and Prepositional Phrase are formed with noun phrases, which can be generalized using a noun concept hierarchy. For example, given a sentence “I met Charlie who is an awesome cat”, we can first extract the event $\langle I, \text{meet, Charlie, -} \rangle$, generalize the event to be $\langle I, \text{meet, \#Cat\#, -} \rangle$, and to more abstract event $\langle I, \text{meet, \#FamilyAnimal\#, -} \rangle$ (assuming we have a concept called “#FamilyAnimal#”). Besides the noun phrases, the predicate phrases can also be generalized. One simple method is to group synonym verbs together as a single concept. For example, “see”, “watch”, and “look” have similar meanings and can be generalized to the “#EyePerceive#” concept and then a more general concept “#PerceptionAction#”. Eventually, the resulting hierarchical knowledge base of

events of this research not only has general category information of each component in an event, but also organizes events in a meaningful way, which can be retrieved based on their abstract concepts.

Besides the research directions presented in the above sections, there are two other research directions that are worth exploring in future. First, we could expand the current affective event knowledge base (the AffectEventKB) by extracting affective events from other data sources. The current AffectEventKB was extracted from the personal story blog posts. However, there are a large amount of other types of texts that can be used for acquiring affective events. For example, it is estimated that hundreds of millions of tweets are tweeted per day. People often discuss events that happened to them in tweets. Therefore, tweets are potentially good data sources for extracting a larger and diverse set of affective events. Comparing and contrasting affective events extracted from different types of texts is also an interesting research topic to explore. In addition, studying and extracting affective events from other languages is another future research direction. It is known that people with different backgrounds are often impacted differently by a same event. For example, eating beef can be an enjoyable experience for many people around the world. However, it could potentially be a taboo for many people with specific religion backgrounds. Automatically acquiring affective events from different languages and understanding their differences are very important for computers to understand how the world works and comprehend natural languages.

Second, it is also a valuable research direction to apply the automatically acquired affective event knowledge base to other applications and demonstrate the effectiveness and usefulness of the knowledge base. Specifically, the learned knowledge base could potentially be useful for fine-grained sentiment understanding, sarcasm detection, plot unit generation, and dialogue response generation. With the knowledge base, the sentiment analysis system can potentially not only predict the overall sentiment polarity of a given text, but also the fine-grained sentiment, i.e., how people are affected by each event described in the text. The knowledge base can also be useful to improve the performance of detecting sarcastic expressions. For example, the expression “I am so happy that mom woke up with vacuuming my room” is sarcastic because the speaker expressed a positive sentiment toward a negative event (i.e., “mom woke up with vacuuming my room”). Prior

polarity of events learned in this research can be very useful to recognize events that are negative based on prior world knowledge. It has been reported that 36% of +/- affect states in fables originate from good or bad situations (i.e., positive or negative events) (Goyal et al., 2013). The affective event knowledge base can potentially help recognize the +/- affect states in fables and then improve the performance of producing plot unit representations. Finally, the prior affective polarity and human need category information of affective events can be useful to help dialogue agents to generate more appropriate responses. For example, in our daily lives, we often talk about the events that happened to us in our conversations with friends and families. When we describe an event to a dialogue agent, the agent needs to understand how the event affects us and why it affects us in that way (i.e., the reason for the event being affective). After understanding these two types of information about the event, the dialogue agent can potentially generate more appropriate responses.

7.3 Summary

This dissertation presents research for learning two types of knowledge about affective events: affective polarity and human needs associated with events. The contribution of this dissertation is twofold. First, this dissertation designed two graph-based semi-supervised models (i.e., Event Context Graph model and Semantic Consistency Graph model) to learn the affective polarity of events. The empirical results demonstrate that many affective events can be identified and assigned prior polarities by these graph-based semi-supervised learning models. The learned affective events (called AffectEventKB) are freely available for the research community. Second, this dissertation proposes a novel research problem of recognizing human needs of affective events, and demonstrates that affective events can be categorized into a small set of human need categories by co-training models with views based on event expressions and event contexts. At the end, this dissertation discusses potential future research directions based on the current research findings.

APPENDIX A

AFFECTIVE POLARITY ANNOTATION GUIDELINES

Different situations affect people in different ways. Situations in this task refer to both dynamic events (e.g., I killed someone) or static states (e.g., I'm happy). For this annotation task, annotators will be given a description of a situation and must decide whether the situation is generally **Desirable**, **Undesirable**, **Mixed**, or **Neutral**.

The judgment should be made from the speaker's perspective (i.e., the person who is describing the situation). Sometimes, the speaker will be explicitly mentioned – in this case, we have replaced instances of “I”, “me”, “we”, and “us” with the general term “@speaker”. If the speaker is not explicitly mentioned, you should assume that the speaker is experiencing the situation (e.g., “bus is late” refers to the speaker's bus, and “mom is sick” refers to the speaker's mom).

A.1 Situation Representation

A situation is represented with four components: **Agent**, **VerbPhrase**, **Object**, and **PrepositionalPhrase**. The Agent is a person or entity performing an action or in a state. The VerbPhrase could be a single verb (e.g., kill) or compound verb phrase (e.g., want to kill). The Object is the object of the verb phrase (i.e., the person or thing that is acted upon). In many cases, one or more of these components will be empty. Empty fields will be represented with a hyphen (-).

All words in the situation representations appear in their root forms. For example, “be” covers all forms of “to be” such as “am”, “is”, and “are”. Similarly, “kill” covers all variations of “to kill” such as “kills” and “killed”.

The following table shows some situation representations, along with examples that illustrate how the situations might appear in actual sentences.

Table A.1: Event Examples and Their Corresponding Sentences.

Situation Representation	Sentences
(@speaker, feel, unappreciated, -)	<p>(1) I am feeling very unappreciated and very underpaid.</p> <p>(2) I felt so ... unappreciated, overwhelmed and overworked.</p> <p>(3) But I just want to say, I feel so unappreciated and all my efforts have gone to waste.</p>
(@speaker, be, - , in disney world)	<p>(1) The world looked so pretty and clean this morning I started to think I was in Disney World.</p> <p>(2) I was in disney world a few years ago and the person i went with looked up and was shocked .</p> <p>(3) I was in the famous Disney World in Orlando Florida with My ex Tony and his sister and brother in law.</p>
(@speaker, be, sick, for time)	<p>(1) But the point is I'm sick for the first time in a while.</p> <p>(2) When I was younger I was really sick for a long time , and that changed everything completely.</p> <p>(3) My parents are freaked out about me being sick so for the first time since.</p>
(@speaker, get, ticket, -)	<p>(1) Then , driving a little too fast , I got a speeding ticket .</p> <p>(2) tonight im gonna go watch mummy3 cuz i got free tickets from work .</p> <p>(3) Bottom line , I got a free ticket from United and was transfered to a Continental flight .</p>
(bus, be, comfortable, -)	<p>(1) The bus was so comfortable .</p> <p>(2) The bus was from RajHamsa Travels , and was comfortable enough for us to fall asleep</p> <p>(3) The Hagey Tour bus was very comfortable.</p>

A.2 Annotation Guidelines

Please assign one of these 4 labels to each situation: **DESIRABLE**, **UNDESIRABLE**, **MIXED**, or **NEUTRAL**.

You will be asked to label each situation generally, in the absence of any specific context, so please try to imagine how you would expect most people to feel when they experience the situation. For example, people typically enjoy going to Disney World so that should be labeled as a Desirable experience, even though of course there will always be exceptions (i.e., a relatively smaller set of people may hate Disney World or have a bad experience while there). No situation will be 100% Desirable, Undesirable, or Neutral. Please focus on the most typical scenarios and use your best judgment as to whether you expect that most people (say > 50%) would consider the situation to be Desirable, Undesirable, or Neutral.

The definitions and guidance for each label are explained below.

- **DESIRABLE**: Use this label if most people would consider the situation to be desirable, enjoyable, pleasant, or beneficial. The speaker should generally be pleased if the situation happens to him/her.

Following are some stereotypical desirable situation:

(@speaker, have, birthday party, -)
 (@speaker, be, - , in disney world)
 (bus, be, comfortable, -)
 (@speaker, find, job, -)

- **UNDESIRABLE**: Use this label if most people would consider the situation to be undesirable, unenjoyable, unpleasant, or detrimental. The speaker should generally be displeased if the situation happens to him/her.

Following are some stereotypical undesirable situations:

(@speaker, be, sick, for time)
 (car , injure, mom, -)
 (bus, be, late, -)
 (@speaker, lose, job, -)

- **NEUTRAL**: Use this label if most people would NOT consider the situation to be desirable or undesirable. For example, many ordinary situations may be neither beneficial nor objectionable.

Following are some stereotypical neutral situations:

(@speaker, walk, - , towards it)

(@speaker, sit, - , -)

(@speaker, drive, - , -)

- **MIX:** Use this label if the situation is rarely neutral but will frequently be considered Desirable by some people and Undesirable by other people. You should feel that both views are common, and it is not clear to you which one is more dominant. For example:

(@speaker, get, ticket, -)

You may imagine common Desirable contexts for this situation (e.g., getting tickets to a concert or sports game) and common Undesirable contexts for this situation (e.g., getting a traffic ticket for speeding or an expired parking meter).

(@speaker, anticipate, snow, -)

You may imagine that some people are very excited about upcoming snowy weather (e.g., skiers and snowboarders) while other people are worried about snowy weather (e.g., people who have travel plans or who hate snow shoveling).

APPENDIX B

EXAMPLE EVENTS WITH AFFECTIVE POLARITY ANNOTATIONS

Table B.1: Examples of Positive Events

(@speaker; trust to do; -; -)	(@speaker's mother; marry; him; -)
(cast; sing; -; -)	(@speaker; kiss; her; -)
(@speaker's apartment; be; clean; -)	(@speakers; want to have; child; -)
(night; be; glad; -)	(@speaker; get; gas money; -)
(-; focus; @speaker's attention; -)	(@speakers; get; girl; -)
(cell phone; have; reception; -)	(day; be; magical; -)
(party; be; okay; -)	(@speakers; spend; hour; at park)
(people; buy; home; -)	(@speaker; get; haircut; -)
(boat ride; be; fun; -)	(@speaker; love; challenge; -)
(@speakers; go; -; to stanley park)	(@speaker; work out; -; -)
(@speaker; work out; 5 day a week; -)	(@speaker's heart; feel; happy; -)
(@speaker; kiss; his chest; -)	(@speaker; kiss; neck; -)
(@speaker; accept; @speaker; -)	(snow; be; bright; -)
(@speakers; win; game; by point)	(@speakers; start to climb; mountain; -)
(@speaker; won; -; -)	(@speakers; be; serious; about it)
(epidural; start to work; -; -)	(walk; be; cool; -)
(lesson; be; good; -)	(@speaker; vanquish; -; -)
(@speaker; want to congratulate; her; -)	(@speaker; play; music; -)
(he; love; @speaker's child; -)	(@speaker's mom; offer to help; -; -)
(@speaker; love; journey; -)	(@speaker; be; foodie; -)
(the day; be; mild; -)	(@speaker; get; compliment; -)
(@speaker; understand; -; in time)	(thing; finish; -; -)
(@speaker; be; glad; for day)	(@speaker; buy; lip gloss; -)
(@speaker; manage to escape; -; -)	(@speakers; get; marry; -)
(@speaker; go; sex; -)	(@speaker; have; your support; -)
(@speaker's family; stay; -; with @speaker)	(@speaker; be; proud; of work)
(one; benefit; -; -)	(@speaker's baby; turn; 4; -)
(-; attract; @speaker; to book)	(@speaker; love; cuteness; -)
(@speaker; go to say; goodnight; -)	(daddy; be; nice; -)
(@speaker; start; -; at @speaker's new job)	(@speaker; seem to accept; -; -)
(@speaker; have; mp3; -)	(palace; be; amazing; -)
(@speaker's birthday; be; amazing; -)	(@speaker; have; choice; -)

Table B.2: Examples of Negative Events

(parking; be; insane; -)	(@speaker; be; busy; with @speaker's homework)
(@speakers; be; bummed; -)	(dog; pass away; -; -)
(life; be; dull; -)	(@speaker; kneel; in front; -)
(@speaker; get; concussion; -)	(@speaker; have to freeze; @speaker's ass; -)
(@speaker; be; -; in cemetery)	(@speaker; not be; good; with thing)
(song; not help; -; -)	(patience; wear; thin; -)
(computer; be; slow; -)	(@speaker; give; grief; -)
(@speakers; give; finger; -)	(@speaker; drop; @speaker's phone; in toilet)
(they; coerce; @speaker; -)	(@speaker's clock; be; wrong; -)
(@speaker; cheat; -; on one)	(people; be; incapable; -)
(thunder; crash; -; -)	(@speaker; avoid; @speaker's blog; -)
(brake; stop to work; -; -)	(people; fuck over; @speaker; -)
(couple; yell; -; -)	(@speaker; not be; happy; at work)
(@speaker; be; scared; about it)	(alice; force; @speaker; -)
(@speaker; come to crash; -; -)	(bath; not help; -; -)
(thing; addicting; -; -)	(@speaker's face; look; pale; -)
(-; lose; @speakers; -)	(@speaker; begin to hyperventilating; -; -)
(@speaker; dump; -; -)	(word; forget; -; -)
(tear; pour; -; from eye)	(-; stick; @speaker; at end)
(house phone; not work; -; -)	(-; swamp; @speaker; with homework)
(@speaker; use to terrify; -; -)	(@speaker; not learn; something; -)
(jerk; say; -; -)	(@speaker; get; -; into argument)
(sucker; be; big; -)	(@speaker; not be; sure; of @speaker's feeling)
(@speaker; clash; -; -)	(@speaker; be; bitch; to everyone)
(@speaker; not eat; dinner; -)	(@speaker; be; busy; in way)
(@speaker; have to wait; while; -)	(boy; be; obsess; -)
(@speaker; feel; burden; -)	(one; be; pushy; -)
(@speaker; be; unloved; -)	(@speaker; crash; -; on your couch)
(engine; stop to run; -; -)	(@speaker; mean; crazy; -)
(computer; be; dead; -)	(-; catch; @speaker; in thunderstorm)
(@speaker; start to snuffle; -; -)	(you; cut down; @speaker; -)
(@speaker; quarrel; -; -)	(god; be; embarrassed; -)
(@speaker; need; toilet; -)	(@speaker; start to blush; -; -)
(@speaker; be; druggy; -)	(he; be; asshole; to @speaker)
(neku; frown; -; -)	(@speaker; need; help; with one)
(@speaker; hear; horror story; -)	(you; annoy; @speaker; -)
(dessert; be to die; -; -)	(you; kill; @speaker; in @speaker's sleep)
(ghost; play; -; -)	(house; be; crowded; -)
(@speaker; refrain; -; -)	(timing; be; difficult; -)
(time; sob; -; -)	(foot; fall; -; -)
(@speakers; wait; -; for aaa)	(-; reduce; @speaker; to mess)
(@speaker; tear; strip; -)	(@speakers; be; low; on gas)
(-; turn off; @speaker's water; -)	(everything; mean; nothing; -)
(you; startle; @speaker; -)	(@speaker's contraction; be; long; -)
(@speakers; be; lower; -)	(stomach; hurt; -; -)

Table B.3: Examples of Neutral Events

(@speaker; send; synopsis; -)	(truth; feel; -; -)
(horoscope; say; -; -)	(@speaker; move out; -; for college)
(way; lay; -; -)	(@speaker; get; call; on tuesday)
(@speakers; run; -; into hiker)	(@speaker; wear; -; around @speaker's neck)
(eye; look; -; -)	(@speaker; have to drink; time; -)
(@speakers; hope to stay; -; -)	(@speaker; try to learn; @speaker's way; -)
(place; have; time; -)	(-; rearrange; room; -)
(world; not do; -; -)	(@speaker; not yell; -; at @speaker)
(@speaker; say; time; to him)	(@speaker; walk; -; into store)
(@speaker; start to take; video; -)	(@speaker; take; his wrist; -)
(@speaker; whole; -; -)	(@speaker; skip; practice; -)
(trunk; be; open; -)	(@speakers; see; color; -)
(home; change; -; -)	(@speaker; have to take; @speaker's dad; -)
(@speaker; stretch out; -; on couch)	(@speakers; get; -; to grandparent house)
(@speaker; go out; way; -)	(@speaker; decide to rent; car; -)
(@speakers; wen; -; -)	(@speaker's contraction; be; -; by time)
(@speaker; want; photo; -)	(@speaker; have; meeting; in morning)
(work; include; -; -)	(weather; call; -; for rain)
(@speakers; decide to buy; them; -)	(@speakers; decide; -; on it)
(@speaker; categorize; -; -)	(she; be; home; with @speaker)
(@speaker; walk; -; to entrance)	(@speaker; flip; -; through them)
(@speaker; think; -; of post)	(@speaker; not have; bra; -)
(@speaker; wake; dream; -)	(@speaker; begin to reflect; -; -)
(@speakers; order; takeout; -)	(@speaker; sit; -; through session)
(@speaker; get; 10 minute; -)	(@speaker; not help; everyone; -)
(nip; be; -; in air)	(@speaker; go; -; to os grid)
(@speaker; scribble; -; -)	(@speakers; have to get; bed; -)
(@speaker; feel to tell; him; -)	(@speaker; download; ringtone; -)
(@speaker; do to keep; her; -)	(@speakers; have to spend; time; -)
(@speaker; have; -; in store)	(@speaker; plan; it; -)
(@speakers; hurry on; -; -)	(ride; slow down; -; -)
(lens don t; have; account; -)	(@speaker; find to challenge; -; -)
(brother; stay; -; -)	(-; build; it; for @speaker)
(@speaker; give; papers; -)	(@speaker; edit out; -; -)
(voice; stop; -; -)	(@speaker; run; -; into @speaker's dad)
(@speakers; drive; -; on freeway)	(type; say; -; -)
(@speaker's cervix; be; open; -)	(@speakers' house; stand; -; -)
(@speaker; go to drag; it; -)	(@speakers; see; cave; -)
(thought; be; much; for @speaker)	(@speaker; go to pick; @speaker's brother; -)
(@speakers; arrive; -; in germany)	(@speakers; end up to watch; rest; -)
(@speaker; have to hang; phone; -)	(tour bus; pull up; -; -)
(neon sign; say; -; -)	(@speaker; hear; @speaker's sister; -)
(@speaker; cought; -; -)	(mommy; bring; @speaker; -)

APPENDIX C

DERIVATION FOR THE SEMANTIC CONSISTENCY GRAPH MODEL

C.1 Computing Update Equations for v and c

In the Semantic Consistency Graph (SCG) model, each event i is associated with a polarity vector v_i , which is a L -dimensional vector where L is the number of polarity labels (i.e., positive, negative, and neutral).

Equation C.1 denotes the objective function of the SCG model, in which v_i and c_k are the parameters to be optimized. Since the objective function is convex, I obtain the update equations by computing the derivatives for these two variables and setting them to zero. Details of the two update equations are describe below.

$$\begin{aligned}
 J(v, c) = & \beta \sum_{i=1}^n D(v_i || v_i^0) + \sum_{(i,j)} \tilde{W}_{ij}^{sim} D(v_i || v_j) + \sum_{(i,j)} \tilde{W}_{ij}^{opp} D(v_i || v_j \mathbf{H}) + \gamma \sum_{(i,k)} \tilde{W}_{ik}^{cmp} D(v_i || c_k) \\
 & + \gamma \sum_{(k,i)} \tilde{W}_{ki}^{cmp'} D(c_k || v_i) + \eta \sum_{k=1}^m D(c_k || c_k^0) \quad (C.1)
 \end{aligned}$$

C.1.1 Update Equation for v_i^{t+1}

In the SCG model, I first compute the update equation for v_i . Given the v^t and c^t , the objective function C.1 can be reformulated to the Equation C.2. The last two terms in C.1 are removed because they are not related to v^{t+1} .

$$\begin{aligned}
 J(v^{t+1}) = & \beta \sum_{i=1}^n D(v_i^{t+1} || v_i^0) + \sum_{(i,j)} \tilde{W}_{ij}^{sim} D(v_i^{t+1} || v_j^t) + \sum_{(i,j)} \tilde{W}_{ij}^{opp} D(v_i^{t+1} || v_j^t \mathbf{H}) \\
 & + \gamma \sum_{(i,k)} \tilde{W}_{ik}^{cmp} D(v_i^{t+1} || c_k^t) \quad (C.2)
 \end{aligned}$$

Then, the partial derivative of the objective C.2 with respect to v_i^{t+1} is computed as the following equation.

$$\begin{aligned} \frac{\partial}{\partial v_i^{t+1}} J(v^{t+1}) &= \beta(\log v_i^{t+1} + 1 - \log v_i^0) + \sum_j \tilde{W}_{ij}^{sim} (\log v_i^{t+1} + 1 - \log v_j^t) \\ &\quad + \sum_j \tilde{W}_{ij}^{opp} (\log v_i^{t+1} + 1 - \log v_j^t \mathbf{H}) + \gamma \sum_k \tilde{W}_{ik}^{cmp} (\log v_i^{t+1} + 1 - c_k^t) \end{aligned} \quad (\text{C.3})$$

Finally, I compute the update equation for v_i^{t+1} by setting the derivative to zero.

$$v_i^{t+1} \propto \exp \frac{1}{O_i} \left(\beta \log v_i^0 + \sum_j \tilde{W}_{ij}^{sim} \log v_j^t + \sum_j \tilde{W}_{ij}^{opp} \log v_j^t \mathbf{H} + \gamma \sum_k \tilde{W}_{ik}^{cmp} \log c_k^t \right) \quad (\text{C.4})$$

where $O_i = \beta + \sum_j \tilde{W}_{ij}^{sim} + \sum_j \tilde{W}_{ij}^{opp} + \gamma \sum_k \tilde{W}_{ik}^{cmp}$.

C.1.2 Update Equation for c_k^{t+1}

Similarly, given v^{t+1} , the original objective function C.1 is reformulated as the Equation C.5.

$$J(c^{t+1}) = \gamma \sum_{(k,i)} \tilde{W}_{ki}^{cmp'} D(c_k^{t+1} || v_i^{t+1}) + \eta \sum_{k=1}^m D(c_k^{t+1} || c_k^0) \quad (\text{C.5})$$

Given v^{t+1} , the partial derivative of the objective function C.5 with respect to c_k^{t+1} is computed as the following equation.

$$\frac{\partial}{\partial c_k^{t+1}} J(c^{t+1}) = \gamma \sum_i \tilde{W}_{ki}^{cmp'} (\log c_k^{t+1} + 1 - \log v_i^{t+1}) + \eta (\log c_k^{t+1} + 1 - c_k^0). \quad (\text{C.6})$$

Finally, I compute the update equation for c_k^{t+1} by setting the derivative to zero.

$$c_k^{t+1} \propto \exp \frac{\eta \log c_k^0 + \gamma \sum_i \tilde{W}_{ki}^{cmp'} \log v_i^{t+1}}{\eta + \gamma \sum_i \tilde{W}_{ki}^{cmp'}} \quad (\text{C.7})$$

C.2 Update Equations for the Component Graph

In Chapter 5, I also designed an independent graph (described in Section 5.2.3.5) for improving component initialization. In this section, I present the derivation process to

compute the update equation for component nodes in the component graph. The following formula is the objective function for the component graph.

$$J_{cmp} = \sum_{(i,j)} \tilde{U}_{ij} D(\mathbf{c}_i | \mathbf{c}_j) + \sum_{i=0}^{m_l} D(\mathbf{c}_i | \mathbf{c}_i^s) + \mu \sum_{i=0}^m D(\mathbf{c}_i | \mathbf{c}_i^0) \quad (\text{C.8})$$

in which m denotes the total number of component nodes, and m_l denotes the total number of components that are contained in the MPQA lexicon.

Given \mathbf{c}^t , the objective function can be reformulated as the following equation,

$$J_{cmp}(\mathbf{c}^{t+1}) = \sum_{(i,j)} \tilde{U}_{ij} D(\mathbf{c}_i^{t+1} | \mathbf{c}_j^t) + \sum_{i=0}^m \delta(i) D(\mathbf{c}_i^{t+1} | \mathbf{c}_i^s) + \mu \sum_{i=0}^m D(\mathbf{c}_i^{t+1} | \mathbf{c}_i^0) \quad (\text{C.9})$$

where $\delta(i) = 1$ if the component i is contained in the MPQA lexicon, otherwise $\delta(i) = 0$. Then, the partial derivative of the above objective function with respect to \mathbf{c}_i^{t+1} can be computed as the following equation.

$$\begin{aligned} \frac{\partial J_{cmp}(\mathbf{c}^{t+1})}{\partial \mathbf{c}_i^{t+1}} &= \sum_j \tilde{U}_{ij} (\log \mathbf{c}_i^{t+1} + 1 - \log \mathbf{c}_j^t) + \delta(i) (\log \mathbf{c}_i^{t+1} + 1 - \log \mathbf{c}_i^s) \\ &\quad + \mu (\log \mathbf{c}_i^{t+1} + 1 - \log \mathbf{c}_i^0) \end{aligned} \quad (\text{C.10})$$

Finally, the update equation for \mathbf{c}_i^{t+1} is computed as the following equation.

$$\mathbf{c}_i^{t+1} \propto \exp \frac{\sum_j \tilde{U}_{ij} \log \mathbf{c}_j^t + \delta(i) \log \mathbf{c}_i^s + \mu \log \mathbf{c}^0}{\sum_j \tilde{U}_{ij} + \delta(i) + \mu} \quad (\text{C.11})$$

APPENDIX D

EXAMPLES OF AUTOMATICALLY LEARNED AFFECTIVE EVENTS

Table D.1: Examples of Automatically Learned Positive Events with Confidence 0.5

(brand; be; best; -)	(band; be; sweet; -)
(@speaker; enjoy; convention; -)	(everything; be; ready; -)
(@speaker; dream; kind; -)	(mountain; be; stunning; -)
(he; inspiring; @speaker; -)	(@speaker; be; careful; in future)
(@speaker; stand to grin; -; -)	(you; be; happy; for @speaker)
(conversation; be; awesome; -)	(activity; be; good; -)
(@speaker; not look; pathetic; -)	(weather; be; wonderful; -)
(procedure; be; successful; -)	(@speaker; love; flickr; -)
(it; be; love; for @speaker)	(@speaker; be; brilliant; -)
(strawberry; be; delicious; -)	(@speaker; want to find; joy; -)
(@speaker; come; -; in peace)	(garden; be; interesting; -)
(@speakers' friend; come to visit; -; -)	(@speaker; get to open; present; -)
(tech; be; nice; -)	(@speaker; enjoy; diving; -)
(family; give; support; -)	(@speakers; be; excellent; -)
(everyone; have; peace; -)	(@speaker; be; grateful; for work)
(@speaker; like; table; -)	(honesty; make; -; -)
(view; be; astounding; -)	(@speakers; have; spending time; -)
(@speaker; love; he; with day)	(technique; be; simple; -)
(@speaker; like; other; -)	(@speaker's new job; be; good; -)
(he; claim to love; @speaker; -)	(night; be; dance; -)
(bit; be; worth; -)	(christmas; be; perfect; -)
(flower; live; -; -)	(@speakers; take; adventure; -)
(@speaker; love to keep; she; -)	(@speaker; love to leave; -; -)
(time; be; good; for @speaker)	(@speaker; hug; -; for time)
(@speaker; love; penguin; -)	(he; pull; @speaker; into kiss)
(@speaker; enjoy; round; -)	(@speaker; like; his hair; -)
(@speaker; thank; father; -)	(@speaker; love; sound; -)
(t; be; nice; -)	(@speaker; like; opera; -)
(@speaker; enjoy; fire; -)	(impressive; know; -; -)
(book; be; amazing; -)	(comfort; be; something; -)
(@speaker; forgiving; -; -)	(god; help out; @speaker; -)

Table D.2: Examples of Automatically Learned Negative Events with Confidence 0.5

(@speaker; smell; trouble; -)	(one; become; sick; -)
(thing; fall; -; through crack)	(@speaker; empty; it; -)
(air conditioner; break; -; -)	(@speaker; bruise up; -; -)
(@speaker's head; fall off; -; -)	(@speaker; be; absurd; -)
(@speaker; feel; stuck; -)	(he; blow; @speaker's mind; -)
(@speaker; itch to know; -; -)	(hallway; be; empty; -)
(@speakers; feel; powerless; -)	(@speaker; have; blister; on @speaker's heel)
(one; seem to bother; -; -)	(-; freak out; @speaker; at time)
(@speaker; go to cut; you; -)	(smile; not go; -; -)
(school; be; hard; -)	(it; not satisfy; @speaker; -)
(-; chastise; @speaker; -)	(@speaker's uncle; be; dead; -)
(two; fall; asleep; -)	(-; go; virus; -)
(@speaker; feel; weak; at knee)	(@speaker; fight; man; -)
(@speaker; lose; 2lb; -)	(-; murder; family; -)
(damage; result; -; -)	(@speaker; find to puzzling; -; -)
(@speaker; waste; day; -)	(@speakers; have; tragedy; -)
(@speaker; be; sarcastic; -)	(@speaker; suffer; -; from anxiety)
(@speaker; feel; bad; for dog)	(@speaker's mind; shut off; -; -)
(battle; go on; -; -)	(@speaker; try to talk; @speaker's way; -)
(@speaker; fight; he; -)	(im; gon to die; -; -)
(@speakers' brain; hurt; -; -)	(battle; be; battle; -)
(problem; take; place; -)	(@speaker; kick; cat; -)
(car; start to slow; -; -)	(@speaker; be; -; like ugh)
(thing; keep to try; -; -)	(@speaker; not want to embarrass; she; -)
(-; confuse; @speaker's mind; -)	(@speaker; not recommend; place; -)
(-; challenge; one; -)	(@speaker; intimidate; people; -)
(people; kill; animal; -)	(step; hurt; -; -)
(guess; be; wrong; -)	(@speaker; mess around; -; with them)
(@speaker; end up to kill; them; -)	(@speaker; miss; son; -)
(idea; be; laughable; -)	(@speaker; not appreciate; him; -)
(@speaker; happen to disagree; -; -)	(@speakers; slow; -; to crawl)
(chest; hurt; -; -)	(@speaker; keep to hurt; people; -)
(-; steal; @speaker's computer; -)	(@speaker's entire body; feel; -; -)
(-; punish; people; -)	(this weekend; be; hard; -)
(@speaker; livid; -; -)	(water; be; filthy; -)
(@speaker; frickin; -; -)	(moan; seem to come; -; -)
(@speaker; roar; -; with laughter)	(@speaker; resent; @speaker; -)
(@speaker; try to scream; time; -)	(it; be; @speaker's ass; -)
(@speakers; be; unlikely; -)	(reason; be; fault; -)
(@speaker; lose; direction; -)	(what; mean; -; -)
(@speaker; keep to lose; weight; -)	(rsquo; feel; bad; -)
(plot; be; similar; -)	(@speaker; not dare to let; -; -)
(@speaker; fall; -; on her bed)	(pier; be; empty; -)
(@speaker; poke; hole; in it)	(@speaker; die; -; of heat exhaustion)
(ride; be; cramped; -)	(hell; be; -; in place)
(one; give; rat ass; -)	(@speaker; split; they; -)
(@speaker; need; knife; -)	(@speaker; fail to remember; -; -)
(work; not be; fun; -)	(@speaker's mouth; feel; -; -)

APPENDIX E

HUMAN NEEDS ANNOTATION GUIDELINES

Affective events are events that typically affect people in positive or negative ways. For this annotation task, we want to know the reason why an event affects someone positively or negatively. You will be given a series of events paired with their *affective polarity* (Positive or Negative). For example: ⟨I, go, , to Disneyland⟩ → Positive. For each event, please determine which of the categories below best explains the reason why the event is Positive or Negative for most people. Use the following statement as a guideline:

The event is Positive because this type of goal has been achieved.
OR
The event is Negative because this type of goal has not been achieved.

In some cases, more than one type of goal may apply. Please choose the one that is **the most relevant reason** to explain the polarity. For example, consider ⟨ , **kidnap, me,** ⟩ → **Negative**. A kidnapping event is Negative with respect to both Physical Safety (category B) and Freedom of Movement (Category G). But for most people, the Physical Safety issue is of greater concern than the Freedom of Movement issue, so this event should be put into Category B. (Please note that in my original guideline, I used the term “Goals”. Later I changed to “Human Needs” which is more appreciate for this research.)

A. PHYSIOLOGICAL GOALS

- (1) **The goal to be able to breathe pleasant or beneficial air, and to avoid unpleasant air.**
- (2) **The goal to avoid hunger, to avoid unpleasant food, and to eat or obtain pleasing food.**
- (3) **The goal to avoid thirst, to avoid unpleasant beverages, and to drink or obtain**

pleasing beverages.

- (4) **The goal to sleep, regularly and comfortably.**
- (5) **The goal to maintain warmth of the human body, to not be too hot or too cold.**
- (6) **The goal to have or obtain shelter (i.e. a place to live or stay) and to avoid unpleasant shelters.**

Examples:

⟨**I, not eat, , for days,** ⟩ → **Negative** is negative because the goal of avoiding hunger is not achieved.

⟨**I, wake up, , at 2am** ⟩ → **Negative** is negative because the speaker failed to achieve the goal of having enough sleep or sleeping soundly.

⟨**I, eat, cake,** ⟩ → **Positive** is positive because the goal of having enjoyable food is achieved.

⟨**I, buy, house,** ⟩ → **Positive** is positive because the goal of owning a shelter is achieved.

B. PHYSICAL HEALTH AND PHYSICAL SAFETY GOALS

- (1) **The goal to be physically healthy and safe.** Affective events in this category could be related to health problems, body injuries, exercise etc.

Examples:

⟨**my head, hurt,** ⟩ → **Negative** is negative because the goal of being physical healthy is failed.

⟨**I, do, yoga,** ⟩ → **Positive** is positive because a goal of yoga is to improve one's health.

⟨ **I, hear, intruder,** ⟩ → **Negative** is negative because the the speaker is probably concerned about their physical safety.

C. LEISURE AND AESTHETIC GOALS

- (1) **The goal to have entertaining (fun) activities, to avoid the lack of fun or entertaining activities.**

- (2) **The goal to have leisure, to avoid too much work because it detracts from leisure time.**
- (3) **The goal to have an enjoyable, pleasant environment**
- (4) **The goal to pursue and appreciate the beauty of nature, art, music and other aesthetically beautiful things.**

Examples:

⟨**I, play, computer game,** ⟩ → **Positive** is positive because it describes a fun activity.

⟨**I, clean, bathroom,** ⟩ → **Negative** is negative because it is typically unenjoyable and detracts from leisure time.

⟨**room, be, noisy,** ⟩ → **Negative** is negative because the environment is undesirable.

⟨**I, see, rainbow,** ⟩ → **Positive** is positive because the goal to pursue/appreciate beauty is achieved.

D. SOCIAL RELATIONS, FAMILY, FRIENDS, SELF-WORTH AND SELF-ESTEEM GOALS

- (1) **The goal to have family, to have close family relations, to avoid damaging family relations.**
- (2) **The goal to have friendships.**
- (3) **The goal to maintain pleasant social relations with others, to avoid conflicts and arguments.**
- (4) **The goal to maintain socially and culturally acceptable behavior.**
- (5) **The goal to realize and improve one's self-worth, to be recognized by others.**
- (6) **The goal to maintain and improve self-esteem or dignity.**

NOTE: if the event describes that emotions/sentiments or opinions are directed to the experiencer or others, then it belongs to this category.

Examples:

⟨**my mom, visit, me,** ⟩ → **Positive** is positive because a family relationship is being maintained.

⟨**I, begin to miss, home,** ⟩ → **Negative** is negative because the goal to be with family

is not satisfied.

⟨**I, have, friend,** ⟩ → **Positive** is positive because the friendship goal is achieved.

⟨**nobody, talk, , to me**⟩ → **Negative** is negative because social relations with others are not good.

⟨**they, mock, me,** ⟩ → **Negative** is negative because the speaker's self-esteem/dignity is hurt.

⟨**, direct, anger, at me**⟩ is negative because bad emotion is directed at me.

E. FINANCES, POSSESSIONS AND JOB SECURITY GOALS

- (1) **The goal to obtain and protect financial income.**
- (2) **The goal to acquire possessions and maintain good condition of one's possessions.**
NOTE: if the possession is more directly related to another type of goal (e.g. food), select that category instead of this one.
- (3) **The goal to have a job and satisfying work.**

Examples:

⟨**I, get, money,** ⟩ → **Positive** is positive because the goal to have financial income is achieved.

⟨**I, buy, computer,** ⟩ → **Positive** is positive because the goal to obtain useful tools is achieved.

⟨**I, lose, my wallet,** ⟩ → **Negative** is negative because the goal to protect possessions is failed.

⟨**I, get fired, ,** ⟩ → **Negative** is negative because the goal to have a job and work well has failed.

F. COGNITION AND EDUCATION GOALS

- (1) **The goal to obtain skills, information, and knowledge, to improve one's intelligence.**
- (2) **The goal to remember and mentally process information correctly.**

Examples:

⟨**I, learn to mow, lawn,** ⟩ → **Positive** is positive because the speaker learned a skill.

⟨**I, graduate, ,** ⟩ → **Positive** is positive because the goal to acquire knowledge is achieved.

⟨**I, forget, ,** ⟩ → **Negative** is negative because the speaker did not process information successfully.

G. FREEDOM OF MOVEMENT AND ACCESSIBILITY GOALS

(1) **The goal to move freely.**

(2) **The goal to access things or services in a timely manner.**

Examples:

⟨**I, wait, , for 5 hours** ⟩ → **Negative** is negative because the goal to access something in a timely manner is not achieved.

⟨**, trap, me, in car** ⟩ → **Negative** is negative because the goal to move freely is not satisfied.

H. MENTAL STATES, EMOTIONS, OR OPINIONS

Use this category label if the event **DOES NOT** belong to previous categories, and meets one of the following rules.

(1) The event directly describes **experiencers' sentiments, emotions, feelings, or physical expressions of emotions.**

(2) The event expresses **some opinions about some objects.**

Examples:

⟨**I, be, happy,** ⟩ → **Positive** because speaker's internal emotion state is positive.

⟨**I, be, mean,** ⟩ → **Negative** because it describes a negative internal mental state.

⟨**Spanish, be, good,** ⟩ → **Positive** because it describes a positive opinion.

IMPORTANT: if an event both expresses a sentiment/emotion and is also related to previous Goal classes, please use the Goal category label. For example, ⟨I, like, cake, ⟩ belongs to the Physiological category.

I. NONE OF ABOVE

Please use this category if:

- (1) **An event or situation is too general or abstract to be assigned to any of the other categories.**
- (2) **The reason why the event is Positive or Negative falls into a different category than the ones listed previously.**

Examples:

⟨**I, have, problem,** ⟩ → **Negative** is negative, but we do not know the specific reason why.

⟨**I, get, mistake,** ⟩ → **Negative** is negative, but we do not know what the mistake is.

REFERENCES

- Andor, D., C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins (2016). Globally Normalized Transition-Based Neural Networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Asur, S. and B. A. Huberman (2010). Predicting the Future with Social Media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence*.
- Baccianella, S., A. Esuli, and F. Sebastiani (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*.
- Banfield, A. (1982). *Unspeakable Sentences*. Routledge & Paul.
- Bennett, K. P. and A. Demiriz (1999). Semi-Supervised Support Vector Machines. In *Advances in Neural Information Processing Systems*.
- Berg-Kirkpatrick, T., D. Burkett, and D. Klein (2012). An Empirical Investigation of Statistical Significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.
- Bermingham, A. and A. F. Smeaton (2011). On Using Twitter to Monitor Political Sentiment and Predict Election Results. *Sentiment Analysis where AI meets Psychology (SAAIP)*, 2.
- Bizer, C., J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann (2009). DBpedia - A Crystallization Point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web* 7(3), 154–165.
- Blum, A. and T. Mitchell (1998). Combining Labeled and Unlabeled Data with Co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. ACM.
- Bollacker, K., C. Evans, P. Paritosh, T. Sturge, and J. Taylor (2008). Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp. 1247–1250. ACM.
- Bruce, R. F. and J. M. Wiebe (1999). Recognizing Subjectivity: A Case Study in Manual Tagging. *Natural Language Engineering* 5(2), 187–205.
- Burton, K., A. Java, and I. Soboroff (2009). The ICWSM 2009 Spinn3r Dataset. In *Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*.
- Burton, K., N. Kasch, and I. Soboroff (2011). The ICWSM 2011 Spinn3r Dataset. In *Proceedings of the Annual Conference on Weblogs and Social Media (ICWSM 2011)*.

- Cambria, E., J. Fu, F. Bisio, and S. Poria (2015). AffectiveSpace 2: Enabling Affective Intuition for Concept-Level Sentiment Analysis. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Cambria, E., D. Olsher, and D. Rajagopal (2014). SenticNet 3: A Common and Common-sense Knowledge Base for Cognition-Driven Sentiment Analysis. In *Twenty-eighth AAAI conference on artificial intelligence*.
- Chang, C.-C. and C.-J. Lin (2011). LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3), 27.
- Chaturvedi, S., D. Goldwasser, and H. Daumé III (2016). Ask, and Shall You Receive? Understanding Desire Fulfillment in Natural Language Text. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.
- Choi, Y., C. Cardie, E. Riloff, and S. Patwardhan (2005). Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*.
- Choi, Y. and J. Wiebe (2014). +/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.
- Chollet, F. et al. (2015). Keras. <https://keras.io>.
- Chung, J. E. and E. Mustafaraj (2011). Can Collective Sentiment Expressed on Twitter Predict Political Elections? In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*.
- Das, D. and S. Petrov (2011). Unsupervised Part-of-Speech Tagging with Bilingual Graph-based Projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.
- Das, D. and N. A. Smith (2011). Semi-supervised Frame-semantic Parsing for Unknown Predicates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.
- Dave, K., S. Lawrence, and D. M. Pennock (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proceedings of the 12th International Conference on World Wide Web*.
- De Marneffe, M.-C. and C. D. Manning (2008). Stanford Typed Dependencies Manual. Technical report.
- Deng, L., Y. Choi, and J. Wiebe (2013). Benefactive/Malefactive Event and Writer Attitude Annotation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.
- Deng, L. and J. Wiebe (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 J. Wiebe (2015a). Joint Prediction for Entity/Event-Level Sentiment Analysis using Probabilistic Soft Logic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Deng, L. and J. Wiebe (2015b). MPQA 3.0: An Entity/Event-Level Sentiment Corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Deng, L., J. Wiebe, and Y. Choi (2014). Joint Inference and Disambiguation of Implicit Sentiments via Implicature Constraints. In *Proceedings of the 25th International Conference on Computational Linguistics*.
- Ding, H. and E. Riloff (2015). Extracting Information about Medication Use from Veterinary Discussions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Ding, H. and E. Riloff (2016). Acquiring Knowledge of Affective Events from Blogs Using Label Propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Ding, H. and E. Riloff (2018a). Human Needs Categorization of Affective Events Using Labeled and Unlabeled Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018)*.
- Ding, H. and E. Riloff (2018b). Weakly Supervised Induction of Affective Events by Optimizing Semantic Consistency. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.
- Dyer, M. G. (1983). The Role of Affect in Narratives. *Cognitive Science* 7(3), 211–242.
- Ekman, P. (1992). An Argument for Basic Emotions. *Cognition & Emotion* 6(3-4), 169–200.
- Esuli, A. and F. Sebastiani (2006). Determining Term Subjectivity and Term Orientation for Opinion Mining. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.
- Fader, A., S. Soderland, and O. Etzioni (2011). Identifying Relations for Open Information Extraction. In *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '11)*.
- Feng, S., J. S. Kang, P. Kuznetsova, and Y. Choi (2013). Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.
- Flanigan, J., C. Dyer, N. A. Smith, and J. Carbonell (2016). CMU at SemEval-2016 task 8: Graph-based AMR Parsing with Infinite Ramp Loss. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1202–1206.
- Fleckenstein, K. S. (1991). Defining Affect in Relation to Cognition: A Response to Susan McLeod. *Journal of Advanced Composition* 11(2), 447–453.
- Go, A., R. Bhayani, and L. Huang (2009). Twitter Sentiment Classification using Distant Supervision. *CS224N Project Report, Stanford*, 1–12.

- Goldberg, A. B., N. Fillmore, D. Andrzejewski, Z. Xu, B. Gibson, and X. Zhu (2009). May All Your Wishes Come True: A Study of Wishes and How to Recognize Them. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.
- Gordon, A. and R. Swanson (2009). Identifying Personal Stories in Millions of Weblog Entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop*.
- Goyal, A., E. Riloff, and H. Daumé III (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., E. Riloff, and H. Daumé III (2013). A Computational Model for Plot Units. *Computational Intelligence* 29(3), 466–488.
- Hochreiter, S. and J. Schmidhuber (1997). Long Short-Term Memory. *Neural computation* 9(8), 1735–1780.
- Höhle, B. (2009). Bootstrapping Mechanisms in First Language Acquisition. *Linguistics* 47(2), 359–382.
- Hu, M. and B. Liu (2004). Mining and Summarizing Customer Reviews. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Jiang, L., M. Yu, M. Zhou, X. Liu, and T. Zhao (2011). Target-dependent Twitter Sentiment Classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.
- Johansson, R. and A. Moschitti (2013). Relational Features in Fine-Grained Opinion Analysis. *Computational Linguistics* 39(3), 473–509.
- Joshi, M., D. Das, K. Gimpel, and N. A. Smith (2010). Movie Reviews and Revenues: An Experiment in Text Regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.
- Jurafsky, D. and J. H. Martin (2016). *Speech and Language Processing*. 3rd ed. draft, November 2016.
- Kang, J. S., S. Feng, L. Akoglu, and Y. Choi (2014). ConnotationWordNet: Learning Connotation over the Word+Sense Network. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.
- Kessler, J. S. and N. Nicolov (2009). Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.
- Kim, S.-M. and E. Hovy (2006). Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*.

- Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.
- Kingma, D. P., S. Mohamed, D. J. Rezende, and M. Welling (2014). Semi-Supervised Learning with Deep Generative Models. In *Advances in Neural Information Processing Systems*.
- Kiritchenko, S. and S. M. Mohammad (2016). The Effect of Negators, Modals, and Degree Adverbs on Sentiment Composition. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*.
- Kiritchenko, S. and S. M. Mohammad (2017). Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2017)*.
- Kiritchenko, S., S. M. Mohammad, and M. Salameh (2016). SemEval-2016 Task 7: Determining Sentiment Intensity of English and Arabic Phrases. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*.
- Kumar, A. and H. Daumé III (2011). A Co-Training Approach for Multi-View Spectral Clustering. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*.
- Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer (2015). DBpedia - A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web* 6, 167–195.
- Lehnert, W. G. (1981). Plot Units and Narrative Summarization. *Cognitive Science* 5(4), 293–331.
- Lenat, D. and R. Guha (1993). Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project. *Artificial Intelligence* 61(1), 4152.
- Lerman, K., S. Blair-Goldensohn, and R. McDonald (2009). Sentiment Summarization: Evaluating and Learning User Preferences. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*.
- Li, J., W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao (2016). Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.
- Li, J., A. Ritter, C. Cardie, and E. Hovy (2014). Major Life Event Extraction from Twitter based on Congratulations/Condolences Speech Acts. In *Proceedings of Empirical Methods in Natural Language Processing*.
- Li, Y. and P. Clark (2015). Answering Elementary Science Questions by Constructing Coherent Scenes using Background Knowledge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.
- Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies* 5(1), 1–167.

- Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.
- Liu, H. and P. Singh (2004). ConceptNet—A Practical Commonsense Reasoning Toolkit. *BT Technology Journal* 22(4), 211–226.
- Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- Maslow, A. H. (1971). The Farther Reaches of Human Nature.
- Maslow, A. H., R. Frager, J. Fadiman, C. McReynolds, and R. Cox (1970). *Motivation and Personality*, Volume 2. Harper & Row New York.
- Mausam, M. Schmitz, R. Bart, S. Soderland, and O. Etzioni (2012). Open Language Learning for Information Extraction. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CONLL)*.
- Max-Neef, M., A. Elizalde, and M. Hopenhayn (1991). *Human Scale Development: Conception, Application and Further Reflections*. The Apex Press.
- McClosky, D., E. Charniak, and M. Johnson (2006). Effective Self-training for Parsing. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*.
- Mendes, P. N., M. Jakob, A. García-Silva, and C. Bizer (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th international conference on semantic systems*, pp. 1–8. ACM.
- Mihalcea, R. (2004). Co-training and Self-training for Word Sense Disambiguation. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*.
- Miller, G. and C. Fellbaum (1998). *Wordnet: An Electronic Lexical Database*. MIT Press Cambridge.
- Mitchell, T., W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling (2015). Never-Ending Learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.
- Mohammad, S., S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry (2016). SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016*.
- Mohammad, S. M. (2018). Word Affect Intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*.
- Mohammad, S. M. and F. Bravo-Marquez (2017). Emotion Intensities in Tweets. In *Proceedings of the Sixth Joint Conference on Lexical and Computational Semantics*.

- Mohammad, S. M. and S. Kiritchenko (2018a). An Annotated Dataset of Emotions Evoked by Art. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*.
- Mohammad, S. M. and S. Kiritchenko (2018b). Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*.
- Mohammad, S. M., S. Kiritchenko, and X. Zhu (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)*.
- Mohammad, S. M., E. Shutova, and P. Turney (2016). Metaphor as a Medium for Emotion: An Empirical Study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*.
- Mohammad, S. M., P. Sobhani, and S. Kiritchenko (2017). Stance and Sentiment in Tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media* 17(3).
- Mohammad, S. M. and P. 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*.
- Mohammad, S. M. and P. D. Turney (2013). Crowdsourcing a Word-Emotion Association Lexicon. 29(3), 436–465.
- Munezero, M. D., C. S. Montero, E. Sutinen, and J. Pajunen (2014). Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. *IEEE Transactions on Affective Computing* 5(2), 101–111.
- Nakov, P., A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov (2016). SemEval-2016 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*.
- Nathanson, D. L. (1994). *Shame and Pride: Affect, Sex, and the Birth of the Self*. WW Norton & Company.
- 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*.
- O'Connor, B., R. Balasubramanyan, B. R. Routledge, and N. A. Smith (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM)*.
- Orbach, M. and K. Crammer (2012). Graph-Based Transduction with Confidence. In *Machine Learning and Knowledge Discovery in Databases - European Conference, (ECML PKDD)*.
- Pang, B. and L. Lee (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2(1-2), 1–135.

- Pang, B., L. Lee, and S. Vaithyanathan (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.
- Pennebaker, J. W., R. J. Booth, and M. E. Francis (2007). Linguistic Inquiry and Word Count: LIWC2007. Austin, TX: *liwc.net*.
- Pennington, J., R. Socher, and C. D. Manning (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Phillips, W. and E. Riloff (2002). Exploiting Strong Syntactic Heuristics and Co-Training to Learn Semantic Lexicons. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*.
- Picard, R. W. (1995). Affective Computing. Technical Report 321, MIT Media Lab Perceptual Computing Technical Report. Revised November 26, 1995.
- Qadir, A. and E. Riloff (2014). Learning Emotion Indicators from Tweets: Hashtags, Hashtag Patterns, and Phrases. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.
- Qadir, A., E. Riloff, and M. Walker (2015). Learning to Recognize Affective Polarity in Similes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Rahimtoroghi, E., J. Wu, R. Wang, P. Anand, and M. A. Walker (2017). Modelling Protagonist Goals and Desires in First-Person Narrative. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*.
- Rao, D. and D. Ravichandran (2009). Semi-supervised Polarity Lexicon Induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*.
- Rashkin, H., S. Singh, and Y. Choi (2016). Connotation Frames: A Data-Driven Investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Rebele, T., F. M. Suchanek, J. Hoffart, J. Biega, E. Kuzey, and G. Weikum (2016). YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. In *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II*.
- Reed, L., J. Wu, S. Oraby, P. Anand, and M. A. Walker (2017). Learning Lexico-Functional Patterns for First-Person Affect. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

- Riloff, E. (1996). Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the National Conference on Artificial Intelligence*.
- Riloff, E., A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang (2013). Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.
- Riloff, E. and J. Wiebe (2003). Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Roberts, K., M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu (2012). EmpaTweet: Annotating and Detecting Emotions on Twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*.
- Russo, I., T. Caselli, and C. Strapparava (2015). SemEval-2015 Task 9: CLIPeval Implicit Polarity of Events. In *Proceedings of the 9th International Workshop on Semantic Evaluation*.
- Sadikov, E., A. G. Parameswaran, P. Venetis, et al. (2009). Blogs as Predictors of Movie Success. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.
- Sang, E. T. K. and J. Bos (2012). Predicting the 2011 Dutch Senate Election Results with Twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*.
- Schank, R. C. and R. P. Abelson (1977). *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. Hillsdale, NJ: L. Erlbaum.
- Scherer, K. R. (2000). Psychological Models of Emotion. *The Neuropsychology of Emotion* 137(3), 137–162.
- Shouse, E. (2005). Feeling, Emotion, Affect. *Media Culture Journal* 8(6).
- Si, J., A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng (2013). Exploiting Topic based Twitter Sentiment for Stock Prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Sobhani, P., S. M. Mohammad, and S. Kiritchenko (2016). Detecting Stance in Tweets And Analyzing its Interaction with Sentiment. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*.
- Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts (2013). Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Speer, R. and C. Havasi (2013). ConceptNet 5: A Large Semantic Network for Relational Knowledge. In *The Peoples Web Meets NLP*, pp. 161–176. Springer.
- Stone, P., D. C. Dunphy, M. S. Smith, and D. M. Ogilvie (1968). The General Inquirer: A Computer Approach to Content Analysis. *Journal of Regional Science* 8(1), 113–116.
- Subramanya, A. and J. Bilmes (2011). Semi-Supervised Learning with Measure Propagation. *Journal of Machine Learning Research* 12, 3311–3370.

- Subramanya, A., S. Petrov, and F. Pereira (2010). Efficient Graph-based Semi-supervised Learning of Structured Tagging Models. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Talukdar, P. P. and K. Crammer (2009). New Regularized Algorithms for Transductive Learning. In *Machine Learning and Knowledge Discovery in Databases, European Conference (ECML PKDD)*.
- Tandon, N., G. de Melo, F. Suchanek, and G. Weikum (2014). Webchild: Harvesting and Organizing Commonsense Knowledge from the Web. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pp. 523–532. ACM.
- Thelen, M. and E. Riloff (2002). A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Theobald, M., J. Siddharth, and A. Paepcke (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*.
- Titov, I. and R. T. McDonald (2008). A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Tokuhsa, R., K. Inui, and Y. Matsumoto (2008). Emotion Classification Using Massive Examples Extracted from the Web. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics.
- Tomkins, S. S. (1962). *Affect Imagery Consciousness: Volume I: The positive Affects*, Volume 1. Springer publishing company.
- Tomkins, S. S. (1963). *Affect, Imagery, Consciousness: Volume II. The Negative Affects.*, Volume 2. Springer publishing company.
- Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*.
- Velikovich, L., S. Blair-Goldensohn, K. Hannan, and R. McDonald (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 (NAACL)*.
- Vu, H. T., G. Neubig, S. Sakti, T. Toda, and S. Nakamura (2014). Acquiring a Dictionary of Emotion-Provoking Events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.
- Wan, X. (2009). Co-Training for Cross-Lingual Sentiment Classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1*.

- Wang, L. and W. Ling (2016). Neural Network-Based Abstract Generation for Opinions and Arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Wiebe, J., T. Wilson, and M. Bell (2001). Identifying Collocations for Recognizing Opinions. In *Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*.
- Wiebe, J., T. Wilson, and C. Cardie (2005). Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* 39(2-3), 165–210.
- Wiebe, J. M. (1994). Tracking Point of View in Narrative. *Computational Linguistics* 20(2), 233–287.
- Wiegand, M., C. Bocionek, and J. Ruppenhofer (2016). Opinion Holder and Target Extraction on Opinion Compounds—A Linguistic Approach. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Wiegand, M. and J. Ruppenhofer (2015). Opinion Holder and Target Extraction based on the Induction of Verbal Categories. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*.
- Wiegand, M., M. Schulder, and J. Ruppenhofer (2016). Separating Actor-View from Speaker-View Opinion Expressions using Linguistic Features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Wilson, T., J. Wiebe, and P. Hoffmann (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*.
- Wu, W., H. Li, H. Wang, and K. Q. Zhu (2012). Probbase: A Probabilistic Taxonomy for Text Understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM.
- Xia, R., C. Wang, X.-Y. Dai, and T. Li (2015). Co-training for Semi-supervised Sentiment Classification Based on Dual-view Bags-of-words Representation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
- Yang, C., K. H. Lin, and H. Chen (2007). Emotion Classification Using Web Blog Corpora. In *Proceedings of the 2007 IEEE / WIC / ACM International Conference on Web Intelligence*.
- Yu, H. and V. Hatzivassiloglou (2003). Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Zhang, L. and B. Liu (2011). Identifying Noun Product Features that Imply Opinions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zhou, D., O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf (2003). Learning with Local and Global Consistency. In *Advances in Neural Information Processing Systems (NIPS)*.

Zhu, X. (2005). *Semi-Supervised Learning with Graphs*. Ph. D. thesis.

Zhu, X. and Z. Ghahramani (2002). Learning from Labeled and Unlabeled Data with Label Propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University.