UNIVERSITY OF CALIFORNIA
SANTA CRUZ

MODELING EVENTS AND AFFECTS IN SOCIAL MEDIA STORIES

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Elahe Rahimtoroghi

March 2018

The Dissertation of Elahe Rahimtoroghi
is approved:

Professor Marilyn Walker, Chair

Professor Pranav Anand

Dr. Hadar Shemtov

Professor Jim Whitehead

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Modeling Events and Affects in Social Media Stories

by

Elahe Rahimtoroghi

Stories play an important role in human perception of the world and therefore the computational analysis of narrative structure is a key area in natural language processing. The focus of this thesis is to develop and evaluate computational models for two main elements of the narrative structure: *Events* and *Desires*. Our work first aims to test a theory that proposes a linear structure of narratives and identifies different parts of a story based on their function. Unlike most of the previous work that use the news articles or other simpler and more conventional genres, we use a corpus of personal stories from social media that have a wider range of topical content and variations of discourse relations. We present an unsupervised method for modeling narrative events, focusing on specific event relations based on the Penn Discourse Treebanks definition of contingency. We use a weakly supervised approach to extract the key events from stories and create a topic-sorted corpus of personal narratives using a bootstrapping method. We additionally propose new evaluation methods for testing the contingent event pairs. Our results show that most of the relations we learn from blog stories are not found in the existing event collections. In our final contribution, we develop supervised methods for modeling the protagonists goals and their outcome in personal narratives, as a sub-problem of modeling affects. Our studies show that both prior and post context are useful for modeling desire fulfillment. In addition, we show that exploiting narrative structure is helpful, both directly in terms of the utility of

discourse relation features and indirectly by using a sequential model. We further examine our analysis of the human desires by identifying and studying the expressions of unfulfilled goals.

For My Parents

Acknowledgments

I would like to express my greatest appreciation to everyone who was a part of my PhD journey. First and foremost, to my advisor, Marilyn Walker, for being an amazing mentor and having the utmost confidence in me. I am tremendously indebted to her for all the inspiring discussions that lead to forming the core ideas of this thesis. This work would not be possible without her guidance, generous support and enthusiasm over the years. Working under the supervision of such a passionate and dedicated professor was an opportunity for which I will always remain grateful.

To Pranav Anand, from whom I have learned massively, starting with the Computational Linguistics course and project. Ever since, his unique ideas and insights have been pivotal in the development of several parts of this thesis, through collaborations and eventually as a member of my dissertation reading committee.

To the rest of my committee, Jim Whitehead and Hadar Shemtov, for taking the time to read my dissertation, their valuable feedback and insightful discussions. I am also greatly thankful to Tracie Tucker, who was the graduate advisor of the Computer Science Department during most of my time as a PhD student, for her continual support and help.

To my dear friends and collaborators in NLDS lab, Stephanie Lukin, Shereen Oraby, Lena Reed, Geetanjali Rakshit, Kevin Bowden, and Vrindavan Harrison, for being there for me any time I needed a friend, each in their own unique way. To Zhichao Hu and Jiaqi Wu, for the excellent work and brilliant ideas in our collaborative projects that shaped parts of this thesis. And Amita Misra, together we started the PhD path and soon we hope to walk the

commencement together.

To my dearest mom, sister and brother for always keeping my heart warm from thousands of miles away, and my dad who was always in my thoughts on this journey. Finally, to my love, Ali, with whom my life's delightful narrative began and yet has many more exciting chapters to unfold.

Chapter 1

Introduction

> *"Narrative is present in myth, legend, fable, tale, novella, epic, history, tragedy, drama, comedy, mime, painting (think of Carpaccio's Saint Ursula), stained glass windows, cinema, comics, news item, conversation. Moreover, under this almost infinite diversity of forms, narrative is present in every age, in every place, in every society; it begins with the very history of mankind and there nowhere is nor has been a people without narrative. All classes, all human groups, have their narratives, enjoyment of which is very often shared by men with different, even opposing, cultural backgrounds. Caring nothing for the division between good and bad literature, narrative is international, trans-historical, trans-cultural: it is simply there, like life itself."*
> — Barthes (1978), Literary Theorist, Philosopher and Linguist

Sharing personal experiences by storytelling is a fundamental aspect of human social behavior (Fivush et al., 2005; Fivush and Nelson, 2004; Bohanek et al., 2008; Habermas and Bluck, 2000; Miller et al., 2007; Bamberg, 2006; Thorne, 2004; McLean and Thorne, 2003; Pratt and Fiese, 2004). More importantly, humans appear to be wired to engage with information that is narratively structured (Gerrig, 1993a; Bamberg, 2006; Bruner, 1991). Telling stories provides a critical developmental and societal function, by serving as a means to reinforce community value systems and to define individual identity (Thorne and Shapiro, 2011; Thorne et al., 2007). This also means that changing people's beliefs through storytelling is easier than

by arguing (Slater and Rouner, 2002). Humans, especially adolescents and young adults who are just establishing their social identity, often tell and retell stories multiple times and multiple ways, seeking the desired response from a peer or community group (Thorne et al., 2004). This has led some theorists to claim that "the stories they tell" is the defining aspect of both individuals and cultures (Bruner, 1991; McAdams et al., 2006; Thorne et al., 2007; Labov and Waletzky, 1967). As stories play an important role in human understanding and perception of the world (Bartlett and Bartlett, 1995), the computational analysis of narrative structure is a key area in natural language processing research.

The purpose of this thesis is to develop computational models of events and desires in everyday lives of humans, expressed in the narratives. Narrative data encompasses a diverse range of texts in the form of a story, such as novels and short stories, poetic and prose epic, film scripts, personal reviews and interviews, oral memoirs, chronicles and histories, and even in the visual forms such as comic strips. What makes a text "narrative" is the sequential structure of events and actions that are connected and evaluated as meaningful consequences. Much of the previous computational work on extracting narrative events and modeling the relations between them is focused on the news genre (Balasubramanian et al., 2013; Chambers and Jurafsky, 2008, 2009; Pichotta and Mooney, 2014; Do et al., 2011). Other work on narratives has developed and evaluated techniques on simpler and more conventional genres such as Aesop's Fables, Wikipedia articles, and simple stories understandable by children (Goyal et al., 2010; Goyal and Riloff, 2013; Chaturvedi et al., 2016).

We believe that research on modeling narrative discourse could highly benefit from using richer datasets with a wider range of topical content and variations of discourse relations.

A summer of rain and dark skies here in Devon - and now, amongst the blustery rainstorms, and occasional bright sunny periods that send me rushing out into the landscape, there is a new crispness in the air, and leaves are falling. Wednesday night, I went with my daughter for a meeting of North Devon arts (NDA), at the Broomhill art hotel, Barnstaple, where we had a delicious and convivial dinner followed by a talk and presentation by photographer, Chris Chapman. Chris is renowned and highly-respected for his documentary photographs of Dartmoor and its farming community during times of crisis and change. The audience was seated in a hushed semi-circle, in the hotel's comfortable lounge/gallery, and we waited expectantly - shifted seats from the centre of the front row to further along the side. My eyes were drawn to the beautiful polished curves of a double bass, leaning against the wall in front of us. And I lingered over its sensuous surfaces and tight strings. And I began to wonder why the hotel should have left a bass and a drum kit lying about... and weren't they just a little bit late in setting up the projection screen for Chris's presentation... I fiddled in my bag, and drew out a notebook and pen. Time to get serious. Then I realised I had left my glasses in the car, and went off into the night to fetch them... and (of course you will have guessed) through the windows of another building, I saw a packed nda meeting in full swing, and the talk about to start. Rushed back, fetched daughter, and explained to the young woman that not only were we in the wrong seats, but the wrong meeting. I was seriously tempted to stay put and listen to the jazz concert - but thank goodness we didn't, because Chris Chapman's talk had a huge impact that I won't easily forget. We found it moving, educational and were riveted by his insights into farming life on Dartmoor. The talk reminded me of the importance of recording and archiving. Things change so fast. We are all aware of this now, more than ever before. It is so easy to forget and lose traditional knowledge and life skills, unless the stories are told and the information cherished...

Figure 1.1: Example of a personal blog story.

We propose that social media is a natural place to find such data, as lots of its content is written by ordinary people about their daily life events and the emotions they experience. Such narratives are rich in common-sense knowledge about humans life and could serve as a valuable resource for developing and evaluating computational models of events and affects. For example, Figure 1.1 shows a fragment of a personal narrative posted on a blog. The concept of personal narrative usually refers to brief and topic-specific stories centered around characters, particular setting and plot. In the personal narratives, the narrators describe and interpret the world and their experiences and draw conclusions about them. They present their inner states and emotions in the form of stories, making them suitable for the goals of our research. There-

fore, we focus our models, methods and evaluations on the personal narratives posted on social media. For example, the story in Figure 1.1 contains descriptions of the **events** that happened to the narrator, such as *"I went with my daughter for a meeting of North Devon arts (NDA)"* and *"we had a delicious and convivial dinner followed by a talk and presentation"*, and the expressions of the narrator's **emotions** and **private states** where some emerge from the events, like *"We found it moving, educational"*, *"The talk reminded me of the importance of recording and archiving."* and *"It is so easy to forget and lose traditional knowledge and life skills"*. It also includes descriptions of the characters in the narrative and the settings of the story, for example the illustration of the weather at the beginning of the narrative and descriptions of the photographer.

Some linguistic theories of narratives posit that human engagement in narrative is partially driven by reasoning about discourse relations between the *events*, and the expectations about what is likely to happen next that results from such reasoning (Gerrig, 1993a; Graesser et al., 1994; Lehnert, 1981; Goyal et al., 2010). The coherence of a story is then provided by the hearer's inferences about causal and motivational relations between story events (Labov and Waletzky, 1967; Lehnert, 1981; Gerrig, 1993b; Goyal et al., 2010). Forster (2010) emphasizes on the importance of causality in a narrative coherence by giving a simple example as follows:

1. *"The king died. Then the queen died."*

2. *"The king died. Then the queen died of grief."*

The first example is only a recitation of a sequence of events that happened in a chronological order. Forster posits that the temporal order is a very primitive feature and it can

only make the hearer want to know what happened next. However, the stories in humans life have additional features. Some events are more important than others, some have emotional impact on people, and some can be related to other events with various relations. For example, the second narrative is a more interesting story where the temporal order is preserved but the sense of causality is stronger. This narrative requires the hearer to remember events and connect them through causal link, and provides information on *why* different incidents happened in the story. The causal relation also connects the characters to each other throughout the story (e.g the king and the queen in this example) and is a more powerful feature than temporal order.

Stories, however, are not just about the events and actions that occur. They are generally situated within a particular time, setting and social group. They also include emotional reactions to those events and their outcomes, either stated or implied (Goyal et al., 2010; Elson and McKeown, 2009, 2010), and often provide "the moral" of the story, which reinforces group value systems. Moreover, sociolinguistic theories of oral narrative, such as Labov & Waletzky's, highlight the social functions of a broad set of narrative elements (Labov and Waletzky, 1967; Labov, 1997; Thorne, 2004; Thorne and McLean, 2003; Thorne et al., 2007). They define three main categories of narrative clauses: temporal, structural, and evaluation points. Structural types include *actions* which report the events, and *orientations* that describe the time and place of the events of the story and identify the participants of the narrative and their initial behavior. For example, the beginning of the story in Figure 1.1 includes orientation clauses that describe the place of the story: *"A summer of rain and dark skies here in Devon"*, *" the blustery rainstorms, and occasional bright sunny periods"*, and *"there is a new crispness in the air, and leaves are falling"*.

According to theories of oral narrative and narrative identity, story events provide a skeleton for conveying the goals and desires of the characters, as well as the affects, attitudes and themes that give the story its point (Fivush et al., 2005; Fivush and Nelson, 2004; Bohanek et al., 2008; Miller et al., 2007; Habermas and Bluck, 2000; Bamberg, 2006; Thorne, 2004; Bohanek et al., 2008; Thorne and McLean, 2003; Pratt and Fiese, 2004). The final element of a story in L&W's categories is *evaluation* point, which they identify as essential to every story. Evaluation gives the reason for telling the story, or the point of it, and without it there is no story, merely a boring recitation of events. The evaluation clauses may also provide information on the consequences of the events as they relate to the goals and desires of the participants, and can be used to describe the events that did not occur, may have occurred, or could occur in the future in the story. The narrative in Figure 1.1 includes examples of evaluation clauses such as *"Chris Chapmans talk had a huge impact that I wont easily forget"*, *"The talk reminded me of the importance of recording and archiving."*, and *"It is so easy to forget and lose traditional knowledge and life skills"*.

Figure 1.2 shows another example of a personal narrative about camping, where some of the narrative elements are highlighted. Theories of narrative generally propose the followings as key elements of a narrative structure:

1. **Static Properties:** The overall setting of the narrative, the participants and their initial behaviors, along with the location and timing of the story. For example, the story in Figure 1.2 includes descriptions of the locations such as *"Sophia's Pizza Restaurant"* and *"American Legion Forest"*.

After eight years, we finally came back to camping ground! Just right before kids ran out of patience, we were there. Weimin is already shaking his head . We were a little bit earlier than checking in time, and then **we went fishing** at a nearby river. **The river is shallow and crystal clear.** You can see the bottom of it. Chinese has a verb: "Clear water has no fish" - don't get me wrong, **we saw a guy right in front of us caught a fish** fly fishing. Many people tubed down too. Once we checked in our camping site, there was a great debate on where to put the tent. As always, the females won. Unfortunately and luckily, the rain started right after we pitched up the tent. **We decided to find a restaurant for dinner,** and went to a local pizza restaurant called **Sophia's Pizza Restaurant**. The pizza there was supper delicious (you must eat the crust too). **It is the best pizza I ever had.** Please make sure you check it out, if you ever come down to **American Legion Forest** at CT. The restaurant is right by Stop Shop, about 3 miles away from our camping site. The rain stopped and ran for the whole two days, but we still managed having fun. We went hiking the second day, and saw thousands of mushrooms. There are more than twenty species of them. Growing up in **Mudanjiang**, where mushroom is very abundant, I love eating mushrooms. **I wish I knew how to identify them**, so that we can bring some back and cook them. Especially, we saw a great pile (about 20 or 40 pounds) of huge mushrooms in the shape and color of chicken crest, but a lot bigger (each ear is about a foot wide, see the photo). **I checked online after coming back and found that it is called Chicken Mushrooms, one of the edibles**. I also satisfied my childish hobby of playing fire. **I set up camp fire** every morning and evening, and **cooked on it**. We had oak smoked chicken, corn bread, BBQ pork spare ribs, homemade chocolate bread, sweet corns, noodles, and homegrown cucumbers and tomatoes. Kids went fishing every day with Dad, though they didn't catch anything, but **enjoyed camping so much, and didn't want to leave.** The next day after we came back from camping was sunny and warm, and **made us look forward to next summer even more!**

Figure 1.2: Example of a blog story about camping.

2. **Events:** The events and actions that create the skeleton for the narrative. The story timeline is then formed by two types of relations between the events:

   (a) **Temporal Order:** The temporal relation between the events that form the the narrative chronology. For example, the two events in the camping narrative, *"we went fishing"* and *"we saw a guy right in front of us caught a fish"*, happen in the chronological order.

   (b) **Causal Links:** The causal relations between the events and actions that give the narrative its coherence and contribute to the story timeline as well. For example, *"I*

set up camp fire" occurs to enable the next event *"cooked on it"*.

3. **Goal Expressions:** The goals and desires of the characters of the story, such as *"We decided to find a restaurant for dinner"* that indicates the goal of the camping group.

4. **Motivational Relations:** The relations between the goals and desires of protagonists and actions they take in the story in order to accomplish them. For example, in Figure 1.2 the narrator expressed his desire *"I wish I knew how to identify them"* and later takes an action motivated by his wish *"I checked online after coming back and found that it is called Chicken Mushrooms, one of the edibles"*.

5. **Evaluations:** The evaluative points of the narrative that provide the consequences of events and indicate the affective states of the characters, including:

   (a) **Private States:** The emotions of the characters in the story and their affective responses to the events which includes the emotional consequences of the events on the characters. For example, the consequence of the camping and fishing experiences are described in the blog story *"enjoyed camping so much, and didn't want to leave"*.

   (b) **Goal Outcomes:** The state of the goals and desires at each point of the story and whether they have been fulfilled or not, such as *"It is the best pizza I ever had"* which indicates the fulfillment of a previously expressed goal (*"We decided to find a restaurant for dinner"*).

   (c) **Hypotheticals:** The events and states that have not occurred in the story but may

have happened or may happen in the future. An example of this type is shown at the end of the story in Figure 1.2 *"made us look forward to next summer even more!"* indicating that another camping trip may happen as a consequence of the events that occurred in this narrative.

There has recently been an upsurge in interest in computational models of narrative structure (Lehnert, 1981; Wilensky, 1982) and story understanding with different lines of work (Rahimtoroghi et al., 2016; Swanson et al., 2014; Ouyang and McKeown, 2015, 2014). The *static properties* of the story, like the characters and locations, can to some extent be identified using current NLP tools. For example, Named Entity Recognition (NER) can be used to identify *"Sophias Pizza Restaurant", "American Legion Forest"* and *"Mudanjiang"* in Figure 1.2 as locations. In addition, other entities such as food (*"pizza", "corn bread"* and *"noodles"*) can be identified using lexical resources and knowledge bases like WordNet (Miller and Fellbaum, 1998) and DBPedia (Auer et al., 2007).

However, there are considerable remaining challenges with modeling the other narrative elements. Previous work on modeling *events* has proposed methods for modeling events and the relations between them (Goyal et al., 2010; Schank et al., 1977; Elson and McKeown, 2009, 2010; Chambers and Jurafsky, 2009; Riaz and Girju, 2010; Beamer and Girju, 2009; Do et al., 2011; Gordon and Swanson, 2009; Manshadi et al., 2008; Gordon et al., 2011). A fundamental task in event modeling is identifying and extracting events from narratives. Events can be described in different ways by nouns, such as *"explosion"* and *"celebration"* and verbs such as *"to eat"* and *"to run"* . However, not all verbs and nouns describe an event, for example the

stative verbs such as *"love"* are used to express emotions and private states and many nouns identify people, places and things. Previous work has tackled this problem to some extent, mainly using lexical resources and corpora annotation (Do et al., 2011; Fauceglia et al., 2015; Riaz and Girju, 2013, 2010), but identifying event-triggering verbs and nouns remains as one of the problems in NLP.

An important part of narrative understanding involves understanding the semantics of events and their relations described in the narrative, such as temporal order and causality. As we discussed earlier, causality is a semantically richer feature and plays a stronger role in story coherence than temporal relations. The causal relation between two events can be expressed both explicitly and implicitly in text. For example, the discourse markers such as *"because"* can indicate explicit causal relations. Additionally, hand-coded lexical and semantic patterns have been used in previous work to identify causal relations in text. For example, Girju (2003) used *Noun-Verb-Noun* template (e.g. *"Noun cause Noun"*) to identify causal pairs like *"mosquitoes cause malaria"*. Overall, much of the previous work on narrative causality has been focused on explicit relations and applies supervised methods using lexico-semantic patterns, knowledge bases, annotated templates and discourse markers (Joskowicz et al., 1989; Kaplan and Berry-Rogghe, 1991; Khoo et al., 2000; Girju, 2003; Do et al., 2011; Riaz and Girju, 2013). Such studies require extensive manual work and their annotations are available on limited sets of narrative data like newswire text. Another limitation of this approach is that other genres like blog stories do not necessarily use explicit markers and patterns to express causal relations and are instead rich in implicit causal information. For example, in the narrative in Figure 1.2, *"setting up the camp fire"* is a pre-condition to enable *"cooking"*, but there is no explicit pattern

or marker to indicate their causal relation. Therefore, new models with minimal supervision are needed for modeling events and their causal relations in narrative genres like blog stories.

There are some studies that use unsupervised techniques for learning the relations between events but not all of them evaluate for causality or any specific event relation (Chambers and Jurafsky, 2008, 2009; Manshadi et al., 2008; Nguyen et al., 2015; Balasubramanian et al., 2013; Pichotta and Mooney, 2014). In addition, most of them are focused on learning of the coarse-grained event relations in newswire genre and only a few use data other than newswire (Hu et al., 2013; Manshadi et al., 2008; Beamer and Girju, 2009). The overwhelming focus on newswire to date has limited the corpora and what can be learned from it, because such data only covers newsworthy topics like *bombing, explosions, war* and *killing* (Chambers and Jurafsky, 2009; Riaz and Girju, 2010; Do et al., 2011). This means that the knowledge available in previous datasets is limited to those types of events, rather than the fine-grained events of everyday experience that may have a contingency link, such as *opening a fridge* enabling *preparing food*, or the event of *getting out of bed* being triggered by *an alarm going off*. In this thesis we aim to learn relations between events from personal narratives in blogs with minimal supervision. We focus on modeling specific event relations, motivated by the definition of *contingency* from PDTB.

In addition to the events and their relations, theoretical models of narrative emphasize on the importance of modeling the effect of events on protagonists' emotions and private states. One theory was proposed as conceptual structures called *Plot Units* for modeling core events and states in the stories and the relationships between them (Lehnert, 1981). The motivation of modeling plot units is that the emotional reactions are central to the notion of a narrative and the

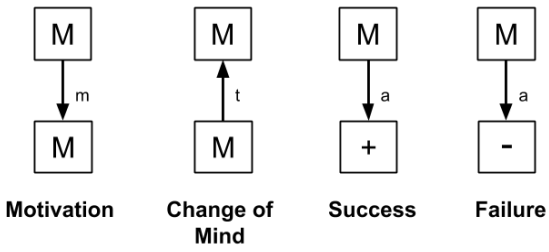

Figure 1.3: Examples of primitive Plot Units

main plot of a story can be modeled by tracking the transition between the affect states (Lehnert, 1981). Figure 1.3 shows the structure of some of the plot units defined by Lehnert. For example, the *Motivation* plot unit consists of a mental state *M* motivating the formation of another mental state, representing a goal or desire. The structure for *Success* consists of a mental state transitioning into a positive state through an actualization link.

There has been limited work on computational models of affects that apply modern NLP tools to identify and track narrative affective states (Chaturvedi et al., 2016; Elson, 2012a; Goyal et al., 2010; Goyal and Riloff, 2013). Previous work presented a system called AESOP that uses a number of existing resources to identify affect states of the characters and create links between them to automatically extract plot units (Goyal et al., 2010; Goyal and Riloff, 2013). This system is evaluated on a small set of two-character stories from Aesop's fables. They identify multiple resources of affects such as direct expression of emotions, situational affect states, and plans and goals and show that most of the affect states emerge by the expression of *goals* and *goal completion*. Their study indicates that current NLP tools are not even sufficient for modeling affects in well-structured and simple two-character stories such as Aesop's fables and identifying affective states requires extensive world knowledge, richer sentiment analysis

resources, and complex inferences about the events and their consequences. In this thesis we aim to model goals and desires of the protagonist in open-domain personal narratives as a step towards modeling a subset of the affect states.

Thesis Outline

The focus of this thesis is to develop and test computational models for two main elements of the narrative structure: *Events and their contingency relation*, and *Goal expressions and their outcomes*. We propose solutions to overcome a subset of the remaining challenges on an open-domain corpus of personal narratives written in blogs. Our methods are presented in three parts as follows:

1. **Learning a Linear Structure of Personal Narratives**

   Personal stories found on the web are sometimes noisy, difficult to interpret and do not always clearly follow the well-defined narrative conventions. Thus, in this thesis we first aim to test a theory that proposes a linear structure of narratives and identifies different parts of a story based on their function. We draw on the theory of oral narrative structure by Labov and Waletzky (1967) that aligns with the nature of the informal blog stories in our corpus. Using this model, we identify three types of clauses in social media narratives based on their purpose in the story:

   - **Orientations:** Clauses that introduce the time and place of the events and identify the participants of the story and their initial behavior.
   - **Actions:** Parts of the story that report the events.

#	Category	Story Clause
1	Orientation	My husband is away.
2	Orientation	This means we live life a little differently than 'everyday'.
3	Orientation	For some reason, being freed of making a 'proper' nightly meal offers me much more time to do things.
4	Evaluation	Perhaps it's the time of day i'm most active ?
5	Evaluation	Who knows.
6	Action	So last night we had cheese and bikkies for dinner
7	Orientation	(it's ok, there was fruit and veg .. okay .. just fruit then, during the day)
8	Action	and watched the ultra-camp "Celebrity Singing Bee"
9	Orientation	(whereas normally, the boy would have to retire to his room
10	Orientation	because it's his bedtime,
11	Orientation	and CSI would be on)
12	Evaluation	God, we had fun.
13	Orientation	It's was 80's night.
14	Action	I romped it in,
15	Evaluation	and i'm sure the boys were impressed with my karaoke skills.
16	Evaluation	Absolutely.
17	Evaluation	I pick my audiences wisely.
18	Evaluation	I hope I don't scar them for life..

Figure 1.4: A personal blog story labeled by L&W's model

- **Evaluations:** Utterances that provide evaluation points of a narrative and information on the consequences of the events as they relate to the goals and desires of the participants.

We created a dataset of personal stories where each clause is labeled by one of these three categories and developed classifiers to automatically identify them in the narratives as shown in the example in Figure 1.4. Our hypothesis was that once this simple struc-

ture is derived from the narrative, it could provide a foundation to identify finer-grained components and infer the relations between them. For instance, we posited that identifying the actions in a narrative could help us infer the causal relations between events. However, as we will discuss in Chapter 4, the utterances in a narrative can have multiple functions so that the categories in Labov's model appear to be too coarse-grained. Our experiments show that identifying these three types does not improve learning of causal relation between events in our corpus.

2. **Modeling Narrative Events and their Relations**

Previous work has mostly focused on learning of coarse-grained event relations from the news articles. They have presented corpora, annotations and extracted event relations, however, their knowledge bases and collections do not provide sufficient coverage of the topics available in the social media narratives. We presents our work on modeling events and the relations between them in Chapter 5 showing that most of the relations we learn from blog stories are not found in the existing event collections (Chambers and Jurafsky, 2009; Balasubramanian et al., 2013).

We focus on specific event relations and use Penn Discourse Treebank (PDTB) definition of *Contingency* which has two types: *Cause* relates two events with a cause-effect relation, and *Condition* relates an event with its possible consequences. For example, in the camping trip story in Figure 1.2, *"seeing a guy catch a fish"* is contingent upon the event expressed before it *"go fishing"*. We present an unsupervised method for learning such contingency relations between events that is tailored to the "oral narrative" nature of blog

stories and directly compare our methods to several other approaches as baselines.

Additionally, we hypothesize that using a topic-specific corpus of narratives where the stories share the same theme could reveal more fine-grained event relations (Riaz and Girju, 2010). We use a weakly supervised approach to extract the key events from stories and create a topic-sorted corpus of personal narratives using a bootstrapping method. We use two different datasets to directly compare what is learned from topic-sorted stories as opposed to a general-domain story corpus. As a part of our model we identify the indicative contingent event pairs from each topic-specific dataset that can potentially be used as building blocks for generating coherent event chains for a particular theme, however, developing the timeline scaffold of the story is outside of the scope of this thesis. In addition, we present new evaluation methods for evaluating contingent event pairs, compared to previous work that mostly uses the Narrative Cloze Test.

3. **Identifying Goal Expressions and their Outcomes**

As we discussed earlier, the study by Goyal and Riloff (2013) indicates that most of the affect states are implicitly expressed in narratives and emerge from the goals and their outcomes. For example, in Figure 1.2 the narrator expresses his desire in the utterance *"I wish I knew how to identify them, so that we can bring some back and cook them"* implying it was not satisfied and resulting in a negative affectual state. Identifying such a desire expression and predicting that it was unfulfilled is a challenging problem that requires deep semantic and syntactic analysis and broad world knowledge. Therefore, our final contribution presented in Chapter 6 is to develop and test methods for recogniz-

ing the expression of the protagonist's goals and desires in narratives and tracking their corresponding outcomes, as a sub-problem of modeling affects. Directly modeling and tracking the affective states and the private states of the protagonist is outside the scope of this thesis.

It has been shown in the previous work that developing computational models of affects, such as learning plot units, even on a small set of well-structured and short stories like Aesop's fables with simplified assumptions, and using manual annotations is still a difficult problem given the performance of the current NLP tools. Consequently, we focus our work on a subset of desire and goals in narratives that are explicitly expressed through a set of verbal patterns.

Other work has attempted to model desire fulfillment in two manually annotated datasets that are deliberately simplified (Chaturvedi et al., 2016). They used *MCTest* which contains 660 stories limited to content understandable by 7-year old children, and *SimpleWiki* created from a dump of the Simple English Wikipedia discarding all the lists, tables and titles. Motivated by this approach, we develop a systematic method for identifying goal expressions in personal narratives and create a new corpus **DesireDB** of desire expressions with their context and annotations of the *desire*, its *fulfillment status*, and the *evidence*. We propose new features as well as testing features used in previous work and apply different classifiers to model desire fulfillment in DesireDB.

Interestingly, our experiments and annotation analysis show that identifying the unfulfillment of a goal is a harder task compared to modeling desire fulfillment, suggesting

that modeling *Thwarted Goals* could be a challenging and novel problem to explore. An example of this is expressed in the first blog story presented earlier in Figure 1.1: *"I was seriously tempted to stay put and listen to the jazz concert - but thank goodness we didn't"*. Here the desire is to *"stay put and listen to the jazz concert"* which is not accomplished. We further examine this idea by creating an additional corpus including $\sim$ 8K expressions of unfulfilled goals. We perform two studies on this dataset. First, we categorize our data to identify common patterns of thwarts, and second, we perform sentiment analysis on the goals and their thwarts to explore whether they can be characterized by the plot unit structure for *Failure*. Our experiments indicate that the sentiment transitions in unfulfilled goals do not always follow the failure plot structure and only about 24% of our data contains thwarts with negative polarity. This result suggests that other conceptual structures are needed to provide a comprehensive model of the goals and their outcomes and also that off-the-shelf sentiment analyzers perform poorly on this genre of narratives.

Overview of Contributions

In conclusion, the contributions of this thesis are as follows:

- Developing a supervised method for automatically identifying a high-level linear structure of the narratives based on Labov and Waletzkys theory of oral narratives. This approach includes creating a dataset with gold-standard labels based on three categories of narrative clauses: Orientation, Action, Evaluation.

- A topic-sorted corpus of personal blog stories, created using a weakly-supervised boot-

strapping method.

- Collections of contingent event pairs, both topic-specific and domain-independent.
- An unsupervised method for learning contingency relation between events from personal narratives. This method includes a semi-supervised algorithm to identify the key events in a topic-specific corpus of narratives. New evaluation methods for testing the contingent event pairs, using human judgments and automatic two-choice questions.
- A corpus of desire expressions, *DesireDB*, with prior and post narrative context including gold-standard annotations for Fulfillment Status, Evidence of Status, and Annotator-Agreement Scores. Construction of this corpus involves a systematic method for identifying and extracting goal expressions in personal narratives.
- Developing supervised methods for modeling the protagonist's goals in general-domain first-person narratives.
- A dataset of *Thwarted Goals* including unaccomplished desire expressions with primary studies on common patterns of thwarts and modeling sentiment transitions.

The rest of this thesis is organized as follows. We discuss the related work in Chapter 2. The corpora and datasets created as a part of this thesis are described in Chapter 3. Our work on learning a linear model of narrative structure based on L&W's theory is presented in Chapter 4. Chapter 5 describes modeling of narrative events and learning contingency relation between them. In Chapter 6 we describe our methods for identifying and modeling desires and goals in first-person narratives. Finally, we present our conclusions in Chapter 7.

Chapter 2

Theoretical Background and Related Work

Linguistics theories have been proposed for modeling narrative discourse, categorizing the main components of a story and defining the relations between them. Such frameworks have been applied in the previous work to derive computational models of narratives. In this Chapter we first present an overview of some of the theoretical frameworks in Section 2.1. A primary purpose of this thesis is to identify events and learning their contingency relations. We present an overview of the previous work on modeling events in narratives in Section 2.2. We discuss supervised and unsupervised methods and the challenges of each approach on different corpora, including datasets like ours. Finally, we survey the previous work on developing computational models of affects in Section 2.3. The previous studies show that current sentiment analysis tools and linguistic resources are not sufficient to reliably identify the affect states and their types. Consequently, in this thesis we focus on a modeling goals and their outcomes as subproblem of modeling affects in narratives. We discuss the previous work on modeling goals and desire fulfillment and compare existing models and datasets to our work.

2.1 Overview of Theories and Frameworks

Previous work on understanding narratives has proposed different approaches for modeling the narrative discourse. We explore three frameworks that motivate the methods presented in this thesis. First, we describe the Theory of Oral Narrative Structure that was proposed by Labov and Waletzky (Labov and Waletzky, 1967; Labov, 1997), and presents a categorization of clauses based on their function in the narrative. In their model, a story must at least have two action clauses with a temporal order, however, they do not investigate the relationships between other clauses and different states in the story. Hence, as the second framework, we explore Plot Units model (Lehnert, 1981) that defines conceptual structures focusing on the links between different states of the characters in the narratives. The emphasis in this model is on the *affect states* and their role in the main plot of a story. The idea is that a story can be modeled by tracking the transition between the affect states of the characters.

However, Plot Units model has limitations. It assumes that all the central events in the story must have either a negative or a positive affect on a character and the affects only emerge from the events, not mental states. In addition, the hypothetical events, such as goals, are not clearly modeled and distinguished from the actual events that happen in the story. As the third framework, we overview the Story Intention Graph (SIG) which presents a richer model for the narrative discourse to overcome the shortcomings of the plot units (Elson, 2012b). The SIG was proposed to capture all the elements in a descriptive representation of the narratives to more robustly express goals, subgoals, plans, and goal outcomes, and the links between the events and the affectual states in the stories.

Theory of Oral Narratives

The theory of oral narratives presented by Labov and Waletzky (L&W) (Labov and Waletzky, 1967; Labov, 1997) divides narrative clauses into three categories: temporal types of narrative clauses, structural clauses, and evaluation points in narratives (Labov, 1997). The *Orientation* and *Action* are two of the structural types of clauses. Labov and Waletzky's model defines an *Action* as a clause reporting an event of the story and define a story as a series of action clauses (events), of which at least two are temporally joined. An *Orientation* clause introduces the time and place of the events of the story, and identifies the participants of the story and their initial behavior. In their model, orientations need to be identified as a separate type of utterance distinct from events to properly understand narrative structure. Figure 2.1 shows an example of a personal narrative written on a blog where the clauses are labeled based on the L&W's theory of oral narrative discourse.

According to Labov and Waletzky, the final element of a story is *Evaluation*, which they identify as essential to every story. Evaluation gives the reason for telling the story, or the point of it and without it there is no story, merely a boring recitation of events. An evaluation clause provides information on the consequences of the events as they relate to the goals and desires of the participants. The evaluation clauses also describe the events that did not occur, may have occurred, or would occur in the future in the story. Thus, a narrative clause in *irrealis* mood (an event or proposition in the narrative that has not actually occurred at, or before, the time the utterance was made as far as the narrator is aware) is an evaluation.

Additionally, L&W define two other structural types: *abstract* and *coda*. Abstract is an initial clause in a narrative that reports the entire sequence of events. A coda is defined

#	Category	Story Clause
1	Abstract	Today was a very eventful work day.
2	Orientation	Today was the start of the G20 summit.
3	Orientation	It happens every year
4	Orientation	and it is where 20 of the leaders of the world come together to talk about how to run their governments effectively and what not.
5	Orientation	Since there are so many leaders coming together their are going to be a lot of people who have different views on how to run the government they follow so they protest.
6	Orientation	This week things started alright and on schedule.
7	Action	There was a protest that happened along the street where I work
8	Action	and at first it looked peaceful until a bunch of people started rebelling
9	Action	and creating a riot.
10	Action	Police cars were burned
11	Action	and things were thrown at cops.
12	Orientation	Police were in full riot gear to alleviate the violence.
13	Action	As things got worse tear gas and bean bag bullets were fired at the rioters
14	Action	while they smash windows of stores.
15	Evaluation	And this all happened right in front of my store
16	Evaluation	which was kind of scary
17	Evaluation	but it was kind of interesting
18	Coda	since I've never seen a riot before.

Figure 2.1: A personal narrative about a protest, labeled by L&W's categories

as a final clause which returns the narrative to the time of speaking, indicating the end of the narrative. The definition for the abstract clauses is ambiguous to some extent; and it is not common to find a single clause reporting all the events in the written narratives. Coda is the last part of an *oral* narrative, when the narrator finishes the story and is as well uncommon in written genres of narratives.

We believe that the narrative structure of the personal stories posted on the blogs will be more similar to oral narrative than they are to classical stories. Therefore in our work, we aim to automatically learn a simple linear narrative structure in the blog stories based on the L&W's theory.

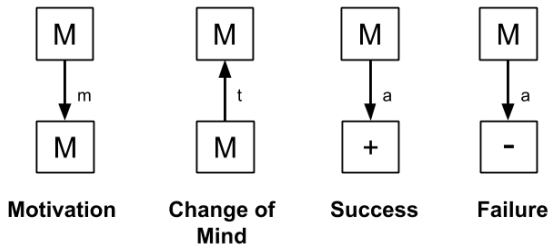

Figure 2.2: Examples of primitive plot units from (Lehnert, 1981)

Plot Units

Plot Units (Lehnert, 1981) were presented as conceptual structures for modeling core events and states in the stories and the relationships between them. The motivation of modeling plot units is that the emotional reactions are central to the notion of a narrative and the main plot of a story can be modeled by tracking the transition between the affect states (Lehnert, 1981). Three types of affect states are defined for modeling Plot Units: Positive (+), Negative (-), and Mental (M); along with two types of links between them: Causal and Cross-Character. Causal links connect the affect states of the same character and have four types: Motivation (m), Actualization (a), Termination (t), and Equivalence (e). Cross-character links connect affect states between different characters, when an event has effect on multiple characters in the story.

Figure 2.2 illustrates four of the primitive plot units with causal links. For example, the transition from a mental state (M) to a negative state (-) by an actualization link (a) is a *Failure* plot unit. The *Change of Mind* plot structure is formed by transition from a mental state to another, with a termination link. The primitive units are the building blocks for other configurations and creating more complex plot units. As an example, the *Nested Subgoals*

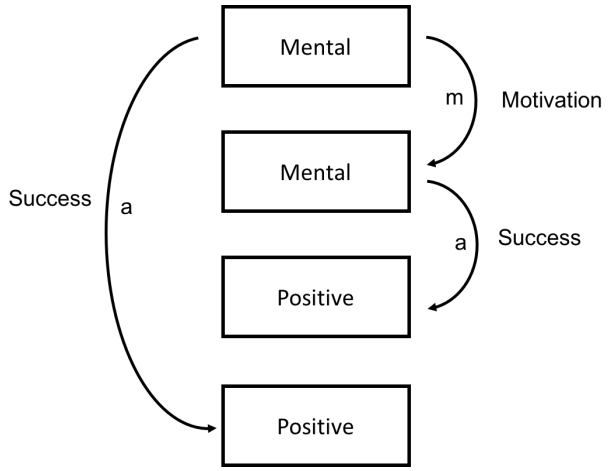

Figure 2.3: Nested Subgoals Plot Unit = Motivation & Success & Success

structure as shown in Figure 2.3, is defined as a combination of a *Motivation* plot unit with two *Success* structures, where both of the mental states in the Motivation end in a positive affect state for the character. This complex plot unit consists of a higher level goal which is represented by the initial mental state, which itself leads to a second mental state as the subgoal. For example, consider the following excerpt from a personal blog story that expresses nested subgoals:

> *I decided on what I wanted to major in. But the college I looked into didn't offer a single thing for my major... so I set out to find a school in Boston that did. I found two.*

The initial goal is wanting to major in a specific field, and the subgoal is to find a school offering it in Boston and both are fulfilled in the story. Figure 2.4 shows the Nested Subgoals structure for this example.

Plot Units were proposed as a representational system for high-level structural analysis in order to generate coherent summaries of the narratives. They form a rich model that analyzes events and represents emotional states of the characters and the relationships between

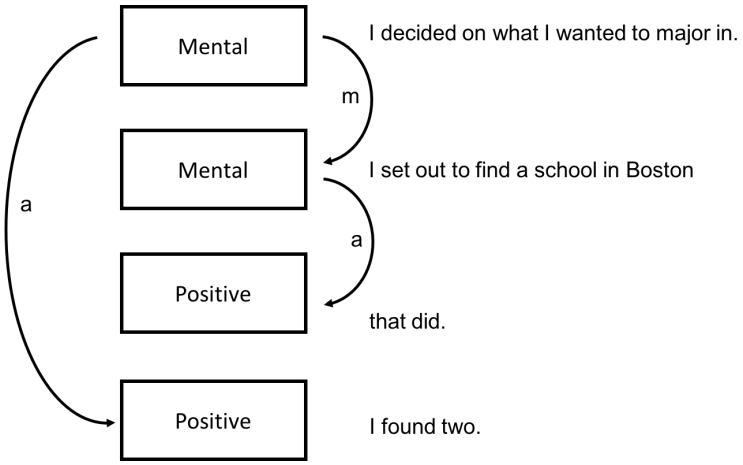

Figure 2.4: An example of nested subgoals expression in a personal narrative

them. However, the way affects are projected on the characters imposes some limitations on the model. In Plot Units structure, it is assumed that the Positive or Negative affect states arise only from the *events* occurring in the story, however, a mental state can also project an affect on a character. For example a desire expression such as *"I wanted to die"* is a mental state that could indicate a negative emotion. In addition, plot units model assumes that every event in the story that is relevant to the plot must necessarily have a negative or a positive affect to an agent. However, it is possible for a character to feel neutral about an event. For example, some of the routine daily activities may not project any positive or negative affect on a person.

Story Intention Graph

A limitation of the Plot Units model is that the distinction between hypothetical states such as goals, and actual events is not clear. The *Story Intention Graph* (SIG) was proposed as a set of discourse relations designed specifically for modeling narrative discourse to overcome

the shortcomings of the plot units (Elson and McKeown, 2009; Elson, 2012b). The idea behind the Story Intention Graph is that the plot units lack the capacity to capture all the elements in a descriptive representation of the narratives and SIG is presented to more robustly express goals, subgoals, plans, beliefs, attempts to achieve goals, and goal outcomes.

The SIG is composed of three layers: Textual layer, Timeline layer, and Interpretative layer. The nodes in the Textual layer contain the original text from the story and together they form the entire narrative. In the Timeline layer the nodes are created based on their temporal order in the story. It represents the chronological order of the events and states of the narrative. The goals, plans, beliefs, and affects are modeled in the Interpretative layer. This layer relates Textual and Timeline layers using their intentional and affectual relations and presents the an agent-oriented interpretation of the narrative. These layers are interconnected subgraphs that together form the Story Intention Graph.

Story Intention Graph is a highly expressive model that is well-motivated by prior models of discourse and narrative comprehension. It encodes the links between an action and the intention of its agent (Poynor and Morris, 2003), between a goal-driven action and its outcome (Magliano and Radvansky, 2001), between a goal and its subgoal or superordinate goal (Richards and Singer, 2001), and between an event and an affectually impacted agent, as well as intersentential relations such as the temporal orderings of events, event modalities (such as suppositions and desires), and links between events and the equivalent source text (such as words or sentences in the original fable). It also presents 12 different *Affect typings* based on prior works on the psychology of human motivation (Elson, 2012c). They are motivations that a character may hold that either motivate or result from actions or events. These motivations are

primarily derived from the work of Maslow (Maslow, 1943) and Max-Neef (Max-Neef et al., 1992) who argue that most human behavior across cultures can be explained by a small finite set of hierarchical motivations. Elson (Elson, 2012c) suggests that such typing can increase the expressiveness of the model.

Though innovative in its emphasis on agency and richness in modeling the narrative discourse, acquiring this level of analysis on user generated content, such as the social media stories, is resource intensive. Such narrative genres do not always clearly follow well defined narrative conventions and are often difficult to interpret and annotate.

2.2 Modeling Narrative Events

Recent work in NLP has tackled the inference of relations between events from a broad range of perspectives: (1) as inference of a discourse relations, e.g. the Penn Discourse Treebank (PDTB) *contingency* relation and its specializations; (2) as a type of common sense reasoning; (3) as part of text understanding to support question-answering; and (4) as way of learning script-like or plot-like knowledge structures (Girju, 2003; Chambers and Jurafsky, 2009; Manshadi et al., 2008; Hu et al., 2013; Nguyen et al., 2015; Balasubramanian et al., 2013; Pichotta and Mooney, 2014). All these lines of work aim to model narrative understanding to enable systems to infer which events are likely to have happened even though they have not been mentioned in the text (Schank et al., 1977), and which events are likely to happen in the future. Such knowledge has practical applications in commonsense reasoning, information retrieval, question answering, narrative understanding and inferring discourse relations.

A fundamental task in event modeling is identifying and extracting the events from narratives. Events can be described in different ways linguistically, mainly through nouns and verbs. Most of the previous work on event modeling define an event as a *predicate-argument* structure and one of the challenges is to extract the verbs and nouns that indicate an event with their arguments. Previous work has studied this problem using lexical resources and corpora annotation (Do et al., 2011; Fauceglia et al., 2015; Riaz and Girju, 2013, 2010). Their results indicate that the major source of event expression is verbs (Fauceglia et al., 2015) and verbs play a very important role in causal relations (Riaz and Girju, 2013). In our work we use a *verb-argument* structure to represent events excluding stative verbs (e.g. *love*) that may not describe an action.

Supervised Identification of Causality

Some previous studies have attempted to extract causal relations from text using knowledge-based inferences (Joskowicz et al., 1989; Kaplan and Berry-Rogghe, 1991). Their work is mainly domain-specific and uses manual annotations and knowledge resources and therefore difficult to scale for many applications. Other work used manually created lexical and semantic patterns to extract causal information (Joskowicz et al., 1989; Kaplan and Berry-Rogghe, 1991; Khoo et al., 2000; Girju, 2003).

Girju (2003) studied different ways that a causal relation can be expressed in the language and showed that the causal expressions can be explicit or implicit in the text. The explicit causal patterns usually contain relevant keywords such as *cause, effect* and *consequence*, however, some of them can be ambiguous like *generate* or *induce* that could imply causal relation

but could also have a different sense. Girju (2003) argues that the implicit causal expressions are more complex and identifying them involves semantic inferences and background knowledge. Therefore, their work focuses on explicit verbal patterns. They created *Noun-Verb-Noun* patterns including a pair of nouns with cause-effect relation, such as *"bonyness"* and *starvation*. Then they used a corpus collected from *LA Times* to extract the verbal patterns that connect such causal pairs of nouns and learned a set of verbs indicating causality like *"induce"* and *"provoke"*. They eventually used a supervised method to learn constraints and classify each *Noun-Verb-Noun* pattern as causal or non-causal.

In a more recent work, two explicit and unambiguous discourse markers, *because* and *but*, were used to automatically collect instances of causal and non-causal event pairs as a training corpus (Riaz and Girju, 2013). They argue that natural disaster and war-related news stories are rich in causal events and thus used 3,000 news articles about "Hurricane Katrina" and the "Iraq war" for creating their event pair corpus. Then they proposed supervised association measures trained on this corpus to learn causality, using lexical, syntactic and semantic features of the event pairs. These features include words, lemmas, part-of-speech tags, all senses from WordNet of the verbs, and verb arguments. They generated three classes of verb pairs: Strongly Causal, Ambiguous, and Strongly Non-causal and evaluated them by two annotators, showing that about 55% of their Strongly Causal verb pairs are judged as causal.

Do et al. (2011) introduced a metric called *Cause-Effect Association* (CEA) which uses Point-wise Mutual Information (PMI) and some components from ECD (proposed previously by Riaz and Girju (2010)) to identify causal relation between events. They first extracted verbal and nominal event predicates with their arguments, using heuristic algorithms and lexical

resources such as WordNet and FrameNet. Then they define an objective function to combine the CEA measure with the discourse relation information, with a set of constraints (e.g. each connective can be only assigned one discourse relation). They hypothesize that using discourse connectives could enhance determining causality between events and their results on a corpus of news articles from CNN shows a 4% increase in F1 while using CEA in combination with discourse classifier as compared to using CEA alone.

The supervised methods like the studies discussed here, require massive manual work and the annotations are available on limited sets of narrative data such as newswire text. In addition, other genres like blog stories do not necessarily use explicit markers and templates to express causal relations and are instead rich in implicit causal information and require new methods for learning causality. One of our primary goals in this thesis is to identify event relations in personal narratives without the use of explicit patterns.

Unsupervised Approaches

There are some studies that use unsupervised techniques for learning event relations but not all of them focus on causality or any specific relation (Chambers and Jurafsky, 2008, 2009; Manshadi et al., 2008; Nguyen et al., 2015; Balasubramanian et al., 2013; Pichotta and Mooney, 2014). Statistical and distributional measures have been used in previous work as unsupervised methods to learn *narrative schema*, *scripts* or *event schema*, which are characterized as collections of events that tend to co-occur (Chambers and Jurafsky, 2008, 2009; Manshadi et al., 2008). Importantly this previous work does not explicitly claim that what is learned is contingent or causal relationships between events, and therefore does not evaluate for causality.

In (Chambers and Jurafsky, 2008, 2009) the Associated Press and New York Times sections of the Gigaword corpus is used to learn narrative schema using Point-wise Mutual Information (PMI). Balasubramanian et al. (2013) generated pairs of relational tuples of events, called *Rel-grams* from a large collection of news wire articles. They use co-occurrence statistics based on Symmetric Conditional Probability (SCP) metric which combines bigram probability considering both directions. The Rel-gram collection is publicly available through an online search interface and is not sorted by topic. They show that their method finds more general associations between relations compared to previous state of the art (Chambers and Jurafsky, 2008, 2009) and has made a first step towards learning event templates at scale. However, their work is focused on news articles and does not consider the causal relation between events for inducing event schema and evaluation.

Riaz and Girju (2010) posit that *contingency* discourse relation plays an important role in language understanding and present an unsupervised method to identify this relation between events. They use scenario-specific news articles on "Hurricane Katrina" and the "Iraq war" to test their model. They hypothesize that the events occurring in a specific scenario tend to have a strong dependency on each other and thus a dataset containing such events is suitable for learning contingency relations. They first cluster sentences into topic-specific scenarios, and then extract scenario-specific events from the sentences in each cluster. They propose two statistical measures: *Effect-Control-Dependency* (ECD) to measure the contingency relation between two events and *Effect-Control-Ratio* (ECR) to identify the Cause (independent event) and the Effect (dependent event). They evaluated their model by two annotators, labeling event pairs as contingent or non-contingent. Their results indicate that using topic-specific dataset

allows to identify strongly contingent event pairs. They also show that distributional measures can be helpful for unsupervised induction of causality.

Another distributional measure, Causal Potential (CP), was introduced by Beamer and Girju (2009) as a way to measure the tendency of an event pair to encode a causal relation, where event pairs with high CP have a higher probability of occurring in a causal context. The causal potential consists of two terms: the first is PMI and the second is relative ordering of bigrams. PMI measures how often events occur as a pair; whereas relative ordering counts how often event order occurs in the bigram. If there is no ordering of events, the relative ordering is zero. This method is unsupervised and focused on causal relations. They generated a corpus of temporally ordered events using movie scripts and showed that CP metric is highly correlated with the probability of two events having a causal relation.

As noted in the the discussions in this Section, there are only few studies which attempt to learn about events and their relationships from data other than newswire (Hu et al., 2013; Manshadi et al., 2008; Beamer and Girju, 2009). The overwhelming focus on newswire to date has limited the available event collections because the news articles only cover newsworthy topics and the knowledge learned is limited to those types of events (Chambers and Jurafsky, 2009; Riaz and Girju, 2010; Do et al., 2011). The more fine-grained events of everyday experience, that may have a contingency link, such as *opening a fridge* enabling *preparing food*, are not covered in these datasets.

In one of our previous studies, we used a corpus of scene descriptions from films and produced initial scalar estimates of potential *contingency* between events using four previously defined measures of distributional co-occurrence (Hu et al., 2013). We showed that none of

the methods from previous work perform better on our data than $\sim 75\%$ average accuracy as measured by human perceptions of contingency. Our results indicated that Causal Potential (CP) achieves a higher accuracy than PMI and Event-Bigram metrics in identifying contingent event pairs in our dataset. In this thesis we build on our previous work and aim to learn contingency relation between events in personal narratives using unsupervised methods. Motivated by the results in (Riaz and Girju, 2010), we generate a topic-sorted corpus of blog stories and show that most of the contingency relations we learn are not found in existing narrative and event schema collections (Chambers and Jurafsky, 2009; Balasubramanian et al., 2013).

2.3 Computational Models of Affects, Desires and Goals

As we discussed in Section 2.1, theories of narrative emphasize the importance of modeling the protagonist's emotions and private states in story understanding. There is broad consensus that understanding a narrative involves activating a representation of the protagonist and her goals and desires and maintaining that representation as the narrative evolves, as a vehicle for explaining the protagonist's actions and tracking narrative outcomes (Elson, 2012c; Rapp and Gerrig, 2006; Trabasso and van den Broek, 1985; Lehnert, 1981). However, there has been limited work on computational models of affects that apply modern NLP tools to identify and track narrative affective states (Chaturvedi et al., 2016; Elson, 2012a; Goyal et al., 2010; Goyal and Riloff, 2013).

A recent work aimed to automatically produce plot unit (Lehnert, 1981) representations for narrative text by categorizing the affect states (Goyal et al., 2010). They created a sys-

tem called AESOP that automatically generates plot units in four steps: affect state recognition, character identification, affect state projection, and link creation. For affect state recognition, they identify words that may be associated with positive, negative, and mental states. They used available resources for sentiment analysis and other type of semantic knowledge bases such as FrameNet, MPQA Lexicon, and Speech Act Verbs. They made simplification assumptions to project the affects to the characters and evaluated their methods on a small set of two-character stories from Aesop's fables. They hand-crafted a simple rule-based coreference resolution system since they use Aesop's fables as the dataset for the experiments and the main characters in the fables are usually animals which makes the coreference resolution challenging.

Then, they map affect states onto characters. Since most plots revolve around events, they use verb argument structure as the primary means for projecting affect states onto characters, using four affect state projection rules. They use Sundance parser (Riloff and Phillips, 2004) to produce a shallow parse of each sentence and simply assume that the subject of the verb phrase is its agent and the direct object is the patient. In all rules, if a clause contains a negation word then the polarity of all words is flipped. Then, simple heuristics are applied for producing links between affect states. They performed a manual annotation to examine different types of affect expressions, indicating that most of the affect states emerge by the expression of goals and plans and goal completion. Their best-performing method achieves an F-measure of 45%, indicating that current natural language processing technology has limitations for plot unit generation. The results of this study shows that modeling affects is a very challenging problem, even on such a simple dataset with manual annotations, several heuristic rules and simplification assumption still achieves a modest performance.

Achieving high performance on modeling of affects and private states requires mature systems for world knowledge, sentiment analysis, reliable affect identification and narrative structure understanding. Overall, the previous work has proven that current sentiment tools and resources are not sufficient for modeling affective states in narratives, specially more informal genres like personal blog stories. One of the challenges in modeling affective states is that people tend to express their emotions and private states implicitly and through the events that happen. Current NLP systems mainly depend on sentiment analysis tools which fail to recognize many events that implicitly impose an affect on the narrative participants. Previous work has studied identifying the affective impact of events on the characters in narratives.

Ding and Riloff (2016) aim to learn a collection of stereotypically positive and negative events from personal blog stories. They created an Event Context Graph where the nodes represent the events and sentences, and the edges connect event mentions in their corpus to their context. They apply a semi-supervised label propagation approach to identify the polarity of each event, resulting on a knowledge base of positive and negative events. They show that many of the events learned by their model can not be correctly labeled by the existing sentiment lexicons.

In a more recent study, Reed et al. (2017) created a dataset of positive and negative first-person sentences from blog stories to learn lexico-functional patterns that reliably predict first-person affect. They show that currently available tools are not sufficient to identify and categorize affects and the performance of current sentiment classifiers can be considerably enhanced by augmenting them with these patterns. Their results indicate the importance of understanding implicit sentiment polarity and that it can represent success or failure of goals and

can be used to better model desire and goal fulfillment in a narrative.

Chaturvedi et al. (2016) exploited two simplified datasets in order to model a desire and its fulfillment status: *MCTest* which contains 660 stories limited to content understandable by 7-year old children, and, *SimpleWiki* created from a dump of the Simple English Wikipedia Richardson et al. (2013); Coster and Kauchak (2011). As most of the affect states emerge from the expression of desires, goals and their outcomes (Goyal and Riloff, 2013), modeling the desires and predicting their fulfillment could serve as a step towards modeling affects. They argue that understanding desire fulfillment requires complex inferences about the expression of desire, events affecting the protagonist (desire subject), and how each part of the story contributes to the completion or unfulfillment of a goal. They used a list of three verb phrases, *wanted to, hoped to,* and *wished to* to identify sentences that include a desire expression and extracted them with context from their datasets. Their context representation consists of five or fewer sentences following the desire expression.

They then extracted features and applied machine learning and statistical techniques to predict whether the desire was fulfilled or not (Chaturvedi et al., 2016). They use BOW (Bag of Words) as baseline and apply unstructured and structured models for desire fulfillment modeling with different features motivated by narrative structure. Their best result is achieved with a structured prediction model called Latent Structured Narrative Model (LSNM) which models the evolution of the narrative by associating a latent variable with each fragment of the context in the data. Their best unstructured model is a Logistic Regression classifier that uses all of their features.

Our work builds on this work by restricting the context around the expression of

desire, but we also examine the role of prior context. The goal of our work is to model the desires of the protagonist in open-domain personal narratives as a step towards modeling a subset of the affect states. We apply deep learning models, and we carry out ablation studies showing which features are more important in predicting goals outcome. In addition we work with naturally occurring social media stories, rather than simpler crowd-sourced datasets.

Chapter 3

Corpora and Annotations

The purpose of this thesis is to develop computational models of events and desires expressed in the personal narratives. As we surveyed in Chapter 2, most of the previous work on computational models of narratives use the newswire corpora (Balasubramanian et al., 2013; Chambers and Jurafsky, 2008, 2009; Pichotta and Mooney, 2014; Do et al., 2011), or other simpler and more conventional genres such as Aesop's Fables, Wikipedia articles, and children stories (Goyal et al., 2010; Goyal and Riloff, 2013; Chaturvedi et al., 2016). We develop and evaluate our methods on a corpus of personal stories from social media that have a wider range of topical content and variations of discourse relations. This Chapter describes our datasets drawn from the Spinn3r corpus of millions of blog posts available as part of the ICWSM 2010 dataset challenge (Burton et al., 2011, 2009).

We first create a subset of the Spinn3r corpus, **Personal Blog Stories**, to include only the personal narratives in the blogs by considering only the posts from six personal blog domains: *livejournal.com*, *wordpress.com*, *blogspot.com*, *spaces.live.com*, *typepad.com*, and *trav-*

Blog story (1):
Things have been kinda crazy lately! We just got back from our cruise and had a fabulous time! I feel rested and relaxed. We returned home to our house basically being finished. We are set to close September 19th. Matt and I both love the house and are feeling very blessed at the moment. I'm finding myself at one of those times in my life where I am waiting on something bad to happen because everything is going so well. I'm not one of those people that can enjoy the good times completely, I still have to worry about something... I think I got that from my granny!

Blog story (2):
I'm not concentrating very well : Science paper today was alright, everyone thought it was hard though, even Liyana (she loves science), But I didn't really think abt whether it was hard or not I just did it I guess? I'm supposed to be studying maths and accounts now but I haven't done anything, naughty girl. Liyana and I went to the central library after our science exam and were planning to study till 5 oh and during our time at the library, Liyana lost her pencil case. It was in her bag together with her wallet and phone but the person only took her pencil case? It was weird but oh wells, so later I'll just pack some stationary into a pencil case for her. At 3 I was really sleepy cause I had only an hour's sleep last night so after that I headed home and reached home at 4 and went straight to sleep, Sheryl climbed into my bed and woke me up at 6.30, to call me for dinner, that cute little one Dinner was pretty special, Aunty Karen said it was a birthday-made dinner, she was actually trying to cook Western food, It was really prettily decorated and yummy thought : thank you. But I didn't manage to finish it cause it was too much of a big portion, I had to pack it up and will bring it to school for lunch tmr, My uncle on the other hand, ate three times of what I ate, he is unbelievable, he kept comparing the amount I ate with the amount my baby cousin eats, Normally she eats more rice but less ingredients, I just eat little rice but lots of ingredients, so I still eat more than her! I don't feel like studying, I want to sleep : But I need to study, ughh : Okay then... I'll go study now.

Figure 3.1: Examples of narratives from Personal Blog Stories corpus

elpod.com . This results in about 800K blog entries. Such narratives are rich in common-sense knowledge about daily life and could serve as a valuable resource for developing and evaluating computational models of narrative events and affects. Figure 3.1 shows two examples from the Personal Blog Stories corpus. The first story is mostly about the narrator's private states, and the expression of feelings and affective states. Whereas the second one includes several event descriptions such as *"Liyana and I went to the central library"*, *"Liyana lost her pencil case"*, and *"My uncle on the other hand, ate three times of what I ate"*. There is also expressions of feelings and private states such as *"he is unbelievable"*, and desires and their consequences like *"I want to sleep : But I need to study, ughh"*. This level of richness and diversity of style,

discourse and topics makes this corpus well-suited for research on computational models of narratives.

From the Personal Blog Stories, we extract a number of datasets targeted for the goals of this thesis. First, to develop a model of narratives based on Labov and Waletzky's theory we create a dataset of personal stories where each clause is labeled by L&W categories. The methods for creating this corpus, the label sets, and annotation process are described in Section 3.1. We present our work on generating topic-sorted sets of personal stories in Section 3.2. We use this corpora for learning fine-grained contingency relation between events. For modeling human desires and goals, we generate datasets of desires expressed in the personal narratives. We describe our main desire expression corpus, *DesireDB*, the data collection and annotation process, along with two other datasets (SimpleDesireDB and ThwartDB) in Section 3.3.

3.1 L&W Dataset

One of the initial goals of our work is to develop a linear structure of the narratives. We draw on the theory of oral narratives by Labov and Waletzky (1967) (L&W) which categorizes the clauses in a story based on their function. This theory provides a linear model of the narratives which aligns well with the nature of the informal narratives in our Personal Blog Stories corpus. We generate annotations for a subset of this corpus to create a dataset for developing and testing methods to identify three types of clauses based on L&W's model: *Orientation*, *Action* and *Evaluation*.

3.1.1 Annotations

We sampled 50 stories from the Personal Blog Stories corpus for annotation and asked annotators to label each clause by one of the three categories *Orientation*, *Action* and *Evaluation*, based on L&W's definitions. As an initial pilot experiment, we previously applied L&W's theory to Aesop's fables and achieved high levels of inter-annotator agreement and high accuracy in machine learning experiments (Rahimtoroghi et al., 2013). However, personal narratives clearly provide a more challenging context for annotation. Therefore, we divided the stories into 4 groups in order to annotate in batches and refined the annotation guidelines after each round based on the disagreements and discussions. We assigned three expert annotators who were familiar with L&W's theory to work on each story. There was a high level of disagreement after the initial batch of annotation, for which we found the following primary sources:

- Clauses of more than one category are common with rising action and evaluation. For example, *"After leaving the apartment at 6:45 AM, flying 2 hours, taking a cab to Seattle, and then driving seven hours up to Whistler including a border crossing, it's safe to say that I felt pretty much like a dick with legs."* which contains elements of orientation, action, and evaluation.

- Actions that are not explicitly stated in the text, but implied in a non-action clause.

- Stative descriptions of the world as a result of an action that are not intuitively orientation.

- Stative descriptions of the world that are localized to a specific place in the narrative, which is problematic to L&W's definition of orientation.

#	Category	Story Clause
1	Orientation	So, unfortunately I couldn't make the Gamesindustry.biz party tonight for two reasons.
2	Orientation	The first was that my phone finally strained out its last ounce of functionality from thelets not extend that metaphorfrom its battery.
3	Orientation	The second was that it takes a really quite unbelievably long time to travel from Brighton to Oxford on trains in the evening;
4	Orientation	just to give you some idea, it involves visiting Gatwick even though youre not getting a plane.
5	Orientation	This did give me a chance to walk aimlessly around the airport shops while waiting for my connection:
6	Evaluation	this is something in which I take a perverse pleasure.
7	Evaluation	I drop the usual thanks to people I met who were nice lest they ever stumble across my postings.
8	Evaluation	I didnt meet anyone who annoyed me even slightly - at a development convention.
9	Action	But I did hear someone say "well, this is certainly sub-optimal" after a fire alarm went off and everyone had to stand around outside.
10	Evaluation	I think the highlight for me was Paul Moore and Tom Johnson's session on movie sound:
11	Evaluation	it really opened my ears to some possibilities by making me think differently about cinematic-style sound design.
12	Evaluation	One of my key problems as an audio engineer is that I tend to make overly-dense and complex mixes -
13	Local-Context	it's a beginner error stemming from insecurity which it takes a lot of people a long time to get over.
14	Stative-Consequence	Listening to movie sound from classics like Raging Bull and Once Upon a Time in the West, you can really hear the benefits of sparsity.
15	Action	On a similar theme, David Moellerstedt from DICE showed their brilliant attenuation system in Battlefield -
16	Local-Context	the game selects key sounds based on perceived loudness and relevance, then ducks the rest of the mix when theyre triggered;
17	Local-Context	it's effectively a side-chaining style effect,
18	Evaluation	and it works so well in a game with such a complex mix.
19	Evaluation	I got a lot of inspiration for Synapse's audio from these talks:
20	Evaluation	it's probably the most productive conference I've ever been to in a lot of ways.
21	Evaluation	Ste Curran should have been awarded something for Best Slide (narrowly beating out a guy from Dolby who had an image of a woman with a toilet-paper-dispensing hat and the headline "Is It Practical?).
22	Evaluation	The reason is simply this:

Figure 3.2: A personal blog story labeled by extended categories

- Subjective clauses that set the scene of the story and have properties of both orientation and evaluation. An example of this is in row 1 of Figure 3.2, *"So, unfortunately I couldnt*

make the Gamesindustry.biz party tonight".

After several rounds of annotation we stabilized on a labeling scheme along with annotation guidelines that annotators could use to disambiguate recurring problematic cases. Our annotation scheme extends the original L&W categories with three additional ones to address the sources of disagreements, as follows:

1. **Local-Context:** A category for distinguishing stative descriptions of the world, that are not intuitively orientation. For example, the clause in row 16 of Figure 3.2 is a stative, describing the game. It is clearly not an action or evaluation, and not intuitively an orientation either, because it is locally dependent.

2. **Implied-Action:** Clauses which do not explicitly mention an action or event, but imply one that is necessary to maintain the causal or temporal coherence of the remaining story. For example rows 4 and 6 in Figure 3.3 imply that the narrator visited the botanical gardens and museums, however the action is not explicitly expressed. In the context of the story it is necessary to know that they did go to those places in order to interpret the other actions described in the narrative. Implied actions are often passive constructions that describe a state of the world that could only be true if an action had taken place.

3. **Stative-Consequence:** A category that describes the state of the world that has resulted as a consequence of an action, but does not directly evaluate the goals, intentions or desires of the participants. An example of this type is row 14 in Figure 3.2, where the narrator describes how you can feel the benefit of sparsity when listening to movie sounds from classics.

#	Category	Story Clause
1	Orientation	Hangzhou quickly became our most familiar city in China as we were drawn there for reasons of proximity, transit, beauty, camera repair, and, most importantly, cheese...
2	Orientation	First up on the culture express was a visit to one of mainstream China's (read: non-Tibetan) most notable temples, the Lingyin Temple...
3	Evaluation	This was the first and only temple experience for the parents in the party, who came away thoroughly impressed and not yet 'templed-out'.
4	Implied-Action	The rest of the afternoon included the botanical gardens, whowere seasonally in a bit of an awkward phase although the presence of magnolias, some lotus flowers, and even a tananger helped justify the ramble in the heat.
5	Action	A quick stroll took us to the NW edge of the lake and the Quyuan Gardens, one of the reputed "New Ten Scenes of West Lake" and home to dozens of species of lotus flowers.
6	Implied-Action	One thing that captured a lot of time in Hangzhou was free or nearly free museums, which Dave good-naturedly accompanied us for despite his bodies insistence to prolong a not so smooth transition to China...
7	Evaluation	you just can't keep the man down...

Figure 3.3: Part of a blog story labeled by extended categories, including Implied Actions

L&W's theory applies to sub-sentence discourse units in a narrative and it is an open question what level of phrasal granularity is appropriate to apply to written narratives. For the annotation process, we treated each independent clause as the basic unit of discourse and manually segmented each story in our dataset using this definition. This resulted in a collection of 1,602 independent clauses. Figures 3.2 and 3.3 show examples of stories from our dataset with annotations.

Using this extended label set we were able to achieve an inter-annotator agreement between the 3 annotators of 0.582 using Fleiss' κ on assigning categories to clauses. We also mapped the full set of labels to a smaller subsets to see if the finer grained distinctions helped improve reliability on more coarse grained labeling schemes. We mapped each extended label to an original L&W category that we thought best fit the original definitions. Local-Context label

Category	Orientation	Action	Evaluation	None-Story
# Clauses	421	436	719	26

Table 3.1: Summary statistics of corpus from L&W label set

was mapped to *Orientation*, Implied-Action to *Action* and Stative-Consequence to *Evaluation*. When mapping back to these reduced label sets we were able to increase the κ to 0.630 for the original scheme with three L&W categories. This result indicates that we can achieve higher reliability by ensuring that the annotators think carefully about particular kinds of distinctions between different stative clauses.

Gold Standard Labels

Gold standard labels were created based on the majority of the annotator assignments. When no annotators agreed on a label, one of the selected labels was chosen randomly. Once completed there were 424 Action clauses, 702 Evaluations, 26 None-stories (clauses in the blog post that were irrelevant to the story), 306 Orientation, 17 Stative-Consequences, 12 Implied-Actions and 115 Local-Contexts. Table 3.1 shows the number of clauses in each category after the three additional labels were mapped to original L&W classes, indicating that *Evaluation* and *Orientation* clauses constitute two thirds of the stories.

3.2 Topic-Sorted Blog Stories

We initially hypothesized that categorizing narrative clauses based on L&W's theory and focusing on *actions* would allow us to improve modeling of events and learning contingency relations between them. However, as we discuss in Chapter 4, using only action clauses did

Camping Trip Narrative 1
That night back at camp, we **made a roaring fire** and **roasted more marshmallows**. We got invited by a few people to join them in their partying, but we stayed in our spot, being old fashioned and cute. That night before bed, we did **take a walk through the rest of the camp site**, and we got invited in to join a group where the men were **singing and passing around a guitar**. After some gentle prodding, **James got up and sang a song** that's very special to us. And I just sat in my chair and melted. The next day, we **made breakfast** amid some mild rainfall, **then ate** said breakfast cozied up under the overhang of our tent. When the rain stopped, **we packed up camp** and headed to Porterville to stay with his parents for the night.
Camping Trip Narrative 2
One day, my friend Do told me she was going to Old Orchard to spend some time by the ocean with two other friends and she told me I should join them. Having nothing in particular to do during the long weekend, I decided to jump in the car with them and go to the States for a little vacation. I am glad I did. **We went camping** (and yes, **we put up the tent** when it was dark) at a campground where we pretty much only spent our evenings and nights there. Every morning we'd **eat breakfast** and **get our picnic ready** and leave for the beach. We'd spend our day **bathing in the sun** and **swam** a little bit - the sea was freezing. I did a lot of reading and **walking**. By the way, I'm reading "The Road" by Cormac McCarthy and I found the book so appropriate for the beach, it's so . sort of soothing and relax (well what I read anyways). In the evening we would **pack up our stuff** and **go back to the campground**, **eat supper** and spend the evening away by the fire camp, **playing** the guitar, **singing**, **eating** wieners and marshmallows.

Figure 3.4: Examples of blog stories from camping trip topic

not improve our results, suggesting that L&W's categories may be too coarse-grained for this purpose. Motivated by previous work (Riaz and Girju, 2010), we posit that using a topic-specific corpus where the stories share the same theme could reveal more fine-grained event relations.

Figure 3.4 shows excerpts of two blog posts about *camping trips* and Figure 3.5 is an example of a general-domain personal narrative (not from any topic-specific subset). Some of the event expressions in these stories are highlighted in bold. We observe that the camping stories include more coherent and strongly related sequences of events, such as *"make a fire and roast marshmallow"*, *"go camping and put up a tent"*, and *"pack up and go back to campground"*. On the other hand, the events expressed in Figure 3.5 are more coarse-grained, like the sequence of events in the last part of the story: *" paused, looked, said"* and *"replied"*.

I went ahead and **bought the camera** after making only $120 so far. I bought the camera online yesterday (because I **got a free memory card** ordering online) And when I went to **look at it** today the price had been reduced $100. AUUGGHH!! So I immediately **called** customer service and was assured that my card would be credited the money back. Crisis and panic adverted - I'm elated!! So now I'm on pins and needles waiting for my camera which should be here sometime next week. Now onto something totally unrelated to the camera but was also something I couldn't resist... I was **standing in line** at work waiting to **pick up a lunch** order. There were a bunch of people down there from different departments of the company - hence a lot of people I don't know. The guy in front of me for example was someone I had never seen before but the guy in front of him works in my lab. So my co-worker and I were talking a little when the guy in between us **looks at me** and does a quick and hard double take. It was so obvious I momentarily **paused** from the conversation and **looked** at him. Probably feeling a little awkward, the guy **said** that I looked exactly like this girl he knew and it through him off for a second. Completely uncharacteristic of me, really I swear - cross my heart, without missing a beat or breaking eye contact I **replied**, "Oh, well she must be really hot". Truly narcissism would be the last trait I would ever use to describe myself but sarcasm may very well be the first and that's the angle was going for here...

Figure 3.5: An example of a general domain blog story

We hypothesize that the narratives that revolve around a particular topic tend to have sequences of more fine-grained events with stronger contingency relation, and therefore the topic-specific datasets could be a better resource for learning contingent event pairs. Consequently, we create topic-specific subsets of the Personal Blog Stories corpus for developing and evaluating our methods presented in Chapter 5, where we focus on learning contingent event pairs.

3.2.1 Corpus Generation

We filter the Personal Blog Stories corpus using a bootstrapping method to generate subsets of it on topics such as *going camping, witnessing a major storm*, and *holiday activities*. We first manually label a small set ($\sim$ 200-300) of stories as our seeds for each topic. Next, we generated the key *event-patterns* from the seed sets using AutoSlog-TS (Riloff, 1996) which is a semi-supervised algorithm. Given hand-labeled stories on a topic and a random set of stories that are not relevant to that topic as input, AutoSlog-TS learns a set of syntactic templates

Topic	Event-Pattern (Caseframe) Examples
Camping Trip	NP-Prep-(NP):CAMPING-IN NP-Prep-(NP):HIKE-TO (subj)-ActVB-Dobj:WENT-CAMPING NP-Prep-(NP):TENT-IN

Table 3.2: Examples of narrative event-patterns learned from Camping Trip stories

(*Caseframes*) that distinguish the linguistic patterns characteristic of the topic from the random set. For each pattern it generates frequency and conditional probability which indicate how strongly the pattern is associated with the topic.

Some examples of such patterns that were learned from camping trip narratives, are provided in Table 3.2. The caseframe patterns were developed for information extraction and search for the syntactic constituent with the designated word as its head. For example, the first row is *NP-Prep-(NP):CAMPING-IN*; This pattern looks for a *Noun Phrase (NP)* followed by a *Preposition (Prep)* where the head of the NP is *"camping"* and the preposition is *"in"*. In our bootstrapping method we use the strong event patterns of each topic (with high frequency and probability) to search the blogs corpus and identify more stories of the same theme. The details of this algorithm and AutoSlog-TS tool are described in Chapter 5.

We used the bootstrapping approach to create topic-sorted datasets of blog stories about different activities. Two examples of *Camping Trip* set were presented in Figure 3.4. Figure 3.6 shows narratives from two other topics: *Visiting a Dentist* and *Holiday Activities* mainly associated with Thanksgiving and Christmas. Our work in Chapter 5 uses two sets of topic-sorted blog stories: *Camping Trip* and *Storm*. The manually labeled dataset includes 361 Storm and 299 Camping Trip stories. After one round of bootstrapping the algorithm identified

Holiday Activities Topic
Well I hope everyone had as wonderful a Christmas as I did. I think it really was the best Christmas I have ever had! We decided not to leave the house and let the family come to us if they wanted. I had lots of fancy bakery treats on hand for breakfast, as well as a huuge pot of coffee. My husband would be cooking a huge lasagna for lunch and I was to cook a giant turkey for dinner with all the fixin's and desserts. We slept til 9 am which is late for us. It's our daughter Audrey's first Christmas so she wasn't gonna wake us up too early. Maybe next year lol. We opened presents over coffee and pastries. The baby ate cheerios... It was just waay too cute watching her attempt to open presents. She was mostly interested in the bows and spent the morning making a collection of them. She got so many toys but her favorite was a large super bouncy ball that I popped in her stocking at the last minute. She loves rolling it around and then chasing it. My husband and I went crazy buying gifts for each other. I got him an MP3 player which he has been wanting for a while now. He got me a great sewing machine which I love. I can't wait to get going on making some baby clothes for summer time and spring!
Visiting a Dentist Topic
You know, I'm not the only one in the family that goes to the dentist. I take my children as well. Well, I take them to their dentist pediatric dentist, but you catch my drift. Well I had Adalia in last week for a bit of dental work. She had a couple of cavities on adult teeth that we wanted to take care of (and hopefully head off life-long dental issues like I have). As we were going over the dental work planned, the assistant pointed to a baby tooth that had a large cavity... touching the adult tooth they were going to fill. She said that had her slated for a stainless-steel crown... but this was on a baby tooth she was supposed to be losing within a year. We discussed the options (crown, leaving it alone, pulling it) and I told her to have the dentist go ahead and pull it. I mean really, a crown on a 12 1/2 year old's baby tooth? When Adalia came back out (slack jawed and drooling) she mentioned to me that the dentist didn't pull her tooth. I went up to the desk and asked the receptionist what happened with that tooth...

Figure 3.6: Examples of topic-specific blog stories

971 additional Storm and 870 more Camping Trip stories. Our experiments and results indicate that using the bootstrapped data considerably improves the accuracy of the contingency model and enhances extracting topic-relevant event knowledge.

3.3 Desire Expression Datasets

The final part of this thesis is focused on modeling the goals and desires of the protagonist and identifying their outcome. For this purpose we create a dataset of desire expressions

People did seem pleased to see me but all I **[wanted to]** do was talk to a particular friend.
I'm off this weekend and had really **[hoped to]** get out and dance.
We **[decided to]** just go for a walk and look at all the sunflowers in the neighborhood.
I **[couldn't wait to]** get out of our cheap and somewhat charming hotel and show James a little bit of Paris.
We drove for just over an hour and **[aimed to]** get to Trinity beach to set up for the night.
She called the pastor, and he had time, too, so, we **[arranged to]** meet Saturday at 9am.
Even though my deadline wasn't until 4 p.m., I **[needed to]** write the story as quickly as possible.

Figure 3.7: Examples of desire expressions in blogs corpus

from the blog stories corpus, **DesireDB**, which includes annotations of the **desire**, its **fulfillment status** and the **evidence** of fulfillment. As we describe in Chapter 6, we perform several feature selection and machine learning experiments on this corpus. Our results indicate that identifying the unfulfillment of a goal is a harder task compared to modeling desire fulfillment. Therefore, we create another dataset, **ThwartDB**, that contains expressions of unfulfilled goals and desires for further analysis of thwarted goals. We describe these two datasets in Sections 3.3.1 and 3.3.2 respectively and present our experiments and analysis on each corpus in Chapter 6.

3.3.1 DesireDB

DesireDB aims to provide a testbed for modeling desire and goals in personal narratives and predicting their fulfillment status. As the first-person narratives often revolve around the narrator's private states and goals (Labov, 1972), the Personal Blog Stories corpus is a highly suitable resource for identifying human desires and their outcomes. Moreover, first-person narratives allow the narrative protagonist to be easily identified and tracked. We develop

a systematic method to identify desire and goal statements in the blog stories, and then collect annotations to create gold-standard labels of fulfillment status as well as spans of text marked as evidence. Figure 3.7 illustrates examples of desire and goal expressions in our corpus.

Identifying Desires and Goals

Human desires and goals can be expressed linguistically in many different ways, including both explicit verbal and nominal markers of desire or necessity (e.g., *want, hope*) and more general markers of urges (e.g., *craving, hunger, thirst*). To systematically discover predicates that specify desires, we browsed FrameNet 1.7 (Baker et al., 1998) selecting frames that seemed likely to contain lexical units specifying desires: *Being-necessary, Desiring, Have-as-a-demand, Needing, Offer, Purpose, Request, Required-event, Scheduling, Seeking, Seeking-to-achieve, Stimulus-focus, Stimulate-emotion,* and *Worry*. We then selected 100 representative instances of that frame in English Gigaword (Parker et al., 2011) by first selecting the 10 most frequent lexical units in that frame, and then selecting 10 random instances per lexical unit. We examined each set of 100 instances, estimating for each sentence whether the predicate specifies a goal that the surrounding text picks up on. Because we were looking for predicates that reliably specify desires that motivate a protagonist's actions, we eliminated frames where less than 80% of the sentences showed this characteristic.

This resulted in a downsample to the following four frames: *Desiring, Needing, Purpose,* and *Request*. We selected only the verbal lexical units because we found that verbs were more likely to introduce goals than nouns or adjectives. We examined 100 instances for each verbal lexical unit, discarding as before. This resulted in 37 verbs. For each verb, we system-

Prior-Context: (1) Anyways, our first stop was up at the Google headquarters in Mountain View. (2) I've heard stories about that place, but damn... none of them really did it justice. (3) Right when we walked in (after we walked past the beach volleyball court), they had a projector showing all the things people were Googleing (Googling?) in real time. (4) It was funny to see all the different things people were looking for – you had your standard argument settling searches (how many home runs did willy Mays hit in 1965), then you had your super obscure searches (what is the national bird of Tansania), and even though the Google people told us that there were a bunch of filters on what shows up on the big board, I caught one search for "milf's"... it's the internet, what're you gonna do? (5) After a bunch of meetings at Google, we wedged ourselves back in the Corolla and headed over toward Santa Cruz.
Desire-Expression-Sentence: I'd never been there, and Adam heard there might be some surf, so we **[wanted to]** check it out.
Post-Context: (1) When we got there, it turned out to be flat as can be, but oh well. (2) We checked out the surf museum in the lighthouse at Steamer Lane, which was pretty tight, then we drove over toward the original o'Neill surf Shop, which was also cool too see. (3) We spent a couple hours there in Santa Cruz, and it was definitely a fun spot... if you get the chance to go there, you should. (4) After we left Santa Cruz, we were planning on driving down through big Sur, but a ridiculous fog rolled in right as we were driving in that direction. (5) I didn't want my first time in big Sur to be ruined by the weather, so we decided to just cruise back to the 101 and high tail it back to Santa Barbara.

Figure 3.8: Example of a desire expression in a personal narrative, with its context

atically constructed and coded all past forms of the verb (e.g., *was [verb]ing*, *had [verb]ed*, *had been [verb]ing*, *[verb]ed*, *didn't [verb]*, etc.) and initially experimented with both past and (historical) present. However, past tense verb patterns resulted in much higher precision. We counted the instances of these patterns in our dataset, and retained only those lemmas with at least 1000 instances across the corpus.

Data Collection

We used the verbal patterns to extract desire expressions from the blogs corpus. To create DesireDB, we extracted desire expressions with five sentences before and after as context. Figure 3.8 shows an example of a desire expression with its narrative context extracted from a blog story. Here, the desire is expressed in *"we **[wanted to]** check it out"* and the narrative context indicates that it has been fulfilled (the first sentence in the post context: *"When*

we got there"). We include both prior and post context of the desire expressions, since theories of narrative structure such as Labov's suggest that the evaluation points of a narrative can precede the expression of the events, goals and desires of the narrator (Labov, 1997). In addition, our annotation results provide support that the evidence of desire fulfillment can be expressed before the desire statement. We also study the effect of prior and post context in understanding desire fulfillment in our experiments and show that using the narrative context preceding the desire statement improves the prediction results.

Data Annotation

We extracted ∼600K desire expressions with their context from the blogs corpus (similar to the example in Figure 3.8). We then sampled 3,680 instances for annotation which consisted of 16 verbal patterns. The annotators were asked to label the fulfillment status of the desire expression sentence based on the prior and post context, by choosing from three labels: *Fulfilled, Unfulfilled,* and *Unknown from the context*. They were also asked to mark the evidence for the label they had chosen by specifying a span of text in the narrative, the subject of the desire expression, and determine if the expressed desire is hypothetical (e.g., a conditional sentence) or not.

The annotators were selected from a list of pre-qualified workers from Amazon Mechanical Turk, who had successfully passed a test on a textual entailment task with 100% correct answers. They were provided with detailed instructions and examples as to how to label the desires and mark the evidence. We also specified the desire expression verbal pattern using square brackets (as shown in Figure 3.8) for more clarity. Three annotators were assigned to work on

Pattern	Count	Fulfilled	Unfulfilled	Unknown	None
wanted to	2,510	49%	35%	14%	2%
needed to	202	65%	16%	16%	3%
ordered	201	71%	21%	6%	2%
arranged to	199	68%	13%	16%	3%
decided to	68	87%	9%	4%	0%
hoped to	68	19%	68%	12%	1%
couldn't wait	68	79%	3%	15%	3%
wished to	66	27%	35%	30%	8%
scheduled	60	43%	25%	27%	5%
asked for	60	53%	27%	15%	5%
required	58	69%	16%	15%	0%
requested	30	60%	20%	20%	0%
demanded	30	60%	23%	17%	0%
ached to	20	50%	40%	10%	0%
aimed to	20	55%	30%	15%	0%
desired to	20	50%	25%	25%	0%
Total	3,680	53%	31%	14%	2%

Table 3.3: Distribution of desire verbal patterns and fulfillment labels in DesireDB

Annotations:
Fulfillment-Label: Fulfilled
Fulfillment-Agreement-Score: 3
Evidence: When we got there, it turned out to be flat as can be, but oh well.
Evidence-Overlap-Score: 3

Figure 3.9: Gold-standard annotations from DesireDB for data instance in Figure 3.8

each data instance. To generate the gold-standard labels we used majority vote and the cases with no agreement were labeled as *"None"*.

Table 3.3 reports the distribution of data and gold-standard labels. About half of the desire expressions (53%) were labeled *Fulfilled* and about one third (31%) were labeled *Unfulfilled*. The annotators didn't agree on about 2% of the instances, that were labeled *None*. As Table 3.3 shows, the distribution of labels is not uniform across different verbal patterns. For instance, *decided to* and *couldn't wait* are highly skewed towards Fulfilled as opposed to *hoped*

Data-Instance:
Prior-Context: Cinderella Lea Salonga is full of awesomeness. It's possible to wish for the impossible. I thought I won't be able to watch Cinderella because of the ridiculous prices (7k for friggin center orchestra seats?!) but the magical world is so good, the other seats were affordable enough.
Desire-Expression-Sentence: I still [wanted to] sit at the orchestra.
Post-Context: Less expensive seats or never get to see Lea on stage? It was a deal, I chose the less expensive seats. We got good seats anyway. I was fidgety until the play started. My insides were soo excited, I kept on checking dad's watch to see if it was 8 pm.

Annotations:
Fulfillment-Label: Unfulfilled
Fulfillment-Agreement-Score: 3
Evidence: I chose the less expensive seats.
Evidence-Overlap-Score: 3

Figure 3.10: A data instance with gold-standard annotations from DesireDB

to which includes 68% Unfulfilled instances. Some patterns seem to be harder to annotate, like *wished to*, which has the highest rate of Unknown (30%) and None (8%) among all.

Other than fulfillment status, for each data instance in our corpus we include the agreement-score which is the number of annotators that agreed on the assigned label. In addition, we provide the *evidence* as a part of the DesireDB data, by merging the text spans marked by the annotators as evidence. We compared the evidence spans pairwise to measure the overlap-score, indicating the number of pairs of annotators with overlapping responses. Figure 3.9 shows the gold-standard annotations in DesireDB for the example data in Figure 3.8.

To assess inter-annotator agreement for Fulfillment, we calculated Krippendorff-alpha Kappa (Krippendorff, 1970, 2004) for pairwise inter-annotator reliability, and, the average of Kappa between each annotator and the majority vote. These two metrics are 0.63 and 0.88 respectively. Overall, 66% of the data was labeled with total agreement (where all three annotators agreed on the same label) and about 32% of data was labeled by two agreements and one disagreement. We also examined the agreements across each label separately. For *Fulfilled*

Fulfilled	Unfulfilled	Unknown	None	Total
1,366	953	380	70	2,780

Table 3.4: Simple-DesireDB dataset

class, total agreement rate is 75%, which for *Unfulfilled* is 67%, and on *Unknown from the context* is 41%. This suggests that annotating unfulfilled desires could be harder than fulfilled cases. For evidence marking, in 79% of the data all three annotators marked overlapping spans.

An example of a data instance with its gold-standard annotations is shown in Figure 3.10. The first part is the extracted data including the desire expression with prior and post context, and the second part is the gold-standard annotations.

Simple-DesireDB

We also constructed a subset of DesireDB in order to be able to compare more directly to the models and data used in previous work. The most related previous work to our research on modeling desires and goals uses three verb phrases to identify desire expressions: *wanted to*, *hoped to*, and *wished to* (Chaturvedi et al., 2016); so we selected a portion of our corpus including these patterns along with two other expressions (*couldn't wait to* and *decided to*) to have sufficient data for experiments. We call this subset *Simple-DesireDB*. Table 3.4 shows the distribution of labels in this dataset.

1	I told Phillip I **wanted to** be there before 4 **but** his haircut took 2 hours... Literally.
2	Trish and I actually **wanted to** abandon the dudes and head off to azure to do some tacky dancing, **but** seems like we can do it as a group.
3	I was a little hesitant to wear fake eyelashes cause I **wanted to** be able to brag that mine were long enough to do without, **but** she put them on me anyway...
4	I desperately **wanted to** answer her back and tell her I did not need extra work load **but** I was not able to sum up the courage too.
5	I **wanted to** apologize to him, **but** I wasn't sure if he would get more mad at me, so I said nothing.
6	We **wanted to** eat Astons, **but** long queue.
7	Malsea **wanted to** go swimming, **but** by the time we got back to the hotel the pool was closed.
8	I **wanted to** kill that sick piece of human waste **but** something touched my heart.
9	I **wanted to** laugh **but** it hurt.
10	I **wanted to** buy a bag, **but** I couldn't find good one.
11	I **wanted to** punch them in the face **but** we sorted out the problem without any face-punching.
12	My body just **wanted to** go back to sleep **but** it didn't work.

Figure 3.11: Examples of thwarted goals, written in blog posts

3.3.2 ThwartDB

ThwartDB is a collection of *thwarted goal* expressions, such as the examples in Figure 3.11. In our experiments we observe that occurrence of the discourse marker *"but"* after the goal expression is a strong indicator of unfulfillment. Motivated by this, we use the pattern ***"wanted to [goal] but [thwart]"*** to extract a collection of thwarted goal expressions from the blog stories.

Examples of sentences matching this pattern are provided in Figure 3.11. In the sentence in row 7 the goal-expression is *"Malsea **wanted to** go swimming,"* and the indication of

unaccomplishment, which we call the thwart-expression, is *"**but** by the time we got back to the hotel the pool was closed"*. The thwart-expression is the clause after *but* that indicates the goal was not fulfilled. We used *wanted to* verbal pattern as our data collection process for DesireDB dataset showed that it was the most frequent pattern in our corpus. We extracted about 8K statements of thwarted goals from the personal blogs. In Chapter 6 we describe our primary studies and experiments on this dataset.

Chapter 4

A Linear Model of Narrative Structure

The purpose of this thesis is to develop computational models of events and desires expressed in personal narratives. Previous work has mostly developed and evaluated their models on conventional narrative genres such as the news articles and short stories. In our research, we use an open-domain corpus of personal narratives written in blogs, which is rich in commonsense knowledge about humans life and could serve as a valuable resource for learning models of events and affects. However, learning narrative models from such a dataset is more challenging than working with simpler corpora. Hence, our initial goal is to test a theory of oral narratives (Labov and Waletzky, 1967) which aligns with the nature of the informal blog stories in our corpus and proposes a linear structure of narratives. We draw on Labov & Waletzky's (**L&W**) model to identify three types of clauses based on their purpose in the story: **(1) Orientation:** clauses that provide traits or properties of the setting or characters, **(2) Action:** clauses that report the events, and **(3) Evaluation**: clauses in the story describing characters' affect states and their emotional reactions to the events.

We choose L&W's theory since the narrative structure of personal blogs would be more similar to oral narrative than classical stories. Also, a deeper analysis and annotation scheme, such as the Story Intention Graph and Plot Units that we described in Chapter 2, is resource intensive on user generated content like blog stories. We hypothesize that once this linear structure is derived from the narrative, it could provide a foundation to identify finer-grained components and infer the relations between them. Particularly, the action clauses in L&W's model indicate the causal relationships between narrative events, and we posit that identifying the actions in a narrative could improve the performance of our methods for modeling events and inferring the causal relations between them.

In this Chapter we present our models and experiments for developing a classifier to automatically identify L&W categories in blog stories. Section 4.1 provides a summary of our annotated dataset for this purpose. We describe our feature extraction and analysis in Section 4.2. Our experiments and results are provided in Section 4.3. As a part of our experimental methods, we present a detailed error analysis on the classifier predictions in Section 4.3.1, and our work for identifying actions to improve modeling narrative events in Section 4.3.2. We summarize our findings and conclusions in Section 4.4.

4.1 L&W Annotated Blogs

As we described in Chapter 3, we created a dataset of 50 blog stories split into clauses where each clause is labeled by one of the three categories *Orientation*, *Action* and *Evaluation*, based on L&W's definitions. After annotation process and assigning gold-standard labels, the

#	Category	Story Clause
1	Orientation	Now, on with this week's story...
2	Orientation	The last month has been hectic.
3	Orientation	Turbo charged.
4	Orientation	Lot's of work because I was learning from Tim, my partner in crime.
5	Orientation	This hasn't been helped by the intense pressure in town due to the political transition coming to an end.
6	Orientation	This week things started alright and on schedule.
7	Action	But I managed to get myself arrested by the traffic police (rouleage) early last Wednesday.
8	Action	After yelling excessively at their outright corrupted methods
9	Action	and asking incessently for what law I actually broke,
10	Action	they managed to bring me in at the police HQ.
11	Action	I was drawing too much of a curious crowd for the authorities.
12	Action	In about half an hour at police HQ I had charmed every one around.
13	Action	I had prepared my "gift" as they wished.
14	Evaluation	Decision withheld, they decided that I neednt to bother,
15	Evaluation	they liked me too much.
16	Evaluation	I should go free.
17	Action	I even managed to meet famous Raus, the big chief.
18	Evaluation	He was too happy to let me go when he realized I was no one.
19	Action	But then, a Major at his side noticed my Visa was expired.
20	Evaluation	Damn!
21	Orientation	My current Visa is being renewed in my other passport at Immigration's.
22	Evaluation	Damn.
23	Evaluation	In custody, for real.

Figure 4.1: An excerpt of a blog story from our corpus

corpus contains 421 Orientations, 436 Actions, 719 Evaluations, and 26 None-stories (clauses in the blog post that were irrelevant to the story). Figure 4.1 shows an example story from our dataset. Our annotation results indicate that action clauses constitute about only one third of the story. Based on this observation, one of our hypotheses is that focusing only on actions would allow us to learn better causal relations between narrative events.

We used the annotated dataset as training and test data for experiments on learning to automatically label narrative clauses. We selected 40 narratives randomly as training and development data and the remaining 10 narratives for the test set. The average story in the training

data had 29.3 clauses with the shortest story consisting of 4 and the maximum consisting of 100. The average story in the test data had 43 clauses with the shortest story consisting of 4 and the maximum consisting of 114. In the rest of this Chapter we describe our feature extraction and classification experiments on this corpus along with the result analysis.

4.2 Feature Representations

To derive feature representations of each type of narrative clause we started with the definitions from the L&W's theory. For instance, the stative verbs such as *"to be"* and *"like"* can indicate the descriptions of orientations and evaluations. Another example is named entities like location and time that can signal orientation clauses. In a primary study, we used such features to develop a classifier on Aesop's fables, achieving a high performance (Rahimtoroghi et al., 2013). We further expanded and refined our primary features by examining L&W's descriptions of distinguishing characteristics of each category.

Since L&W state that the unit of analysis should be an independent clause with its subordinate clauses, we used the Stanford Parser to distinguish independent and dependent clauses and kept track of the features that occurred in both types of clause separately. Distinguishing between the features occurring in the two clause types would allow us to determine if and when the features of the subordinate clause were relevant or more informative for automatic classification.

We defined three sets of features, as shown in Table 4.1. For the Linguistic features, we identified the part-of-speech of the main verb (POS) within both dependent and indepen-

Feature Set	**Description**
Linguistic	Parts of Speech, Dependency parse, Unigrams, Bigrams
Lexical and Sentiment	LIWC counts and frequencies, Negation
Story Position	First-Clause, Last-Clause, Bin-Position (based on binning into ten story regions)

Table 4.1: Feature sets for clause type classification

Feature	InfoGain	Action	Orientation	Evaluation
POS:IND-VBD	0.128	**0.084**	0.002	0.031
BinPosition0	0.076	0.017	**0.042**	0.014
STEM:IND-be	0.074	**0.045**	0.022	0.003
FirstClause	0.044	0.010	**0.019**	0.011
POS:IND-VBZ	0.042	**0.029**	0.008	0.002
IND-Negate	0.040	0.025	0.000	**0.013**
IND-Copula	0.039	**0.030**	0.004	0.005
IND-FirstPerson	0.035	**0.017**	0.004	0.002
IND-LIWC-Motion	0.034	**0.021**	0.003	0.006
POS:IND-VBP	0.033	**0.023**	0.001	0.007
BinPosition9	0.031	0.010	0.007	**0.014**

Table 4.2: The 10 most highly correlated features with each label and cumulatively over all the labels using mutual information and information gain

dent clauses, as well as dependency relations (DEP), lexical unigrams (STEM), and bigrams. We additionally identified whether each clause contained a negation (Negate) and extracted lexical semantic categories from LIWC (Pennebaker et al., 2001). We also developed a set of features describing the relative position of the clause in the story (Bin-Position, First-Clause, Last-Clause), because different story regions are often associated with different clause types. For example, in Figure 4.1 the beginning of the story contains more orientation clauses, while action clauses are concentrated in the middle of the story. The evaluation clauses typically occur part-way through the story where they provide the narrator's reaction to story events.

#	Category	Story Clause with Highlighted Features
1	Orientation	It **was (STEM:IND-be)** the end of the line for the Interpolers. **(BinPosition0, FirstClause)**
2	Orientation	Their last show **was (STEM:IND-be, IND-Copula)** on Friday, the first day of the festival.
3	Action	**I (IND-FirstPerson) walked (POS:IND-VBD, IND-LIWC-Motion)** up on stage on Sunday to find this familiar little artifact, Interpol's keyboards.
4	Action	**I (IND-FirstPerson)** sent Ally an email with a picture of it.
5	Action	He responded rather quickly suggesting that I just push it into a ditch and leave it there for future civilizations to ponder over.
6	Action	Now I'm back in my living room, digging my home and SF.
7	Orientation	My neighbor, incidentally, **is (IND-Copula, STEM:IND-be)** an extra from Mad Max,
8	Evaluation	and I think that he **is (STEM:IND-be)** awesome.
9	Evaluation	I have this elaborate fantasy that involves eating pizza with him on a merry-go-round. **(BinPosition8)**

Figure 4.2: A story from L&W annotated dataset, illustrating some extracted features in bold

In total there were 3,510 unique binary valued features extracted from our training dataset. We used mutual information to find the features that had the highest correlation with each category and the information gain over all the labels. Some of the highest valued features are presented in Table 4.2. The top feature is (POS:IND-VBD) which means the part-of-speech (POS) of the main verb in the independent clause (IND) is past tense (VBD). Figure 4.2 shows an example of a blog story from L&W annotated corpus where some of the top extracted features are highlighted in bold. Row 1 is an orientation clause which contains three features: *STEM:IND-be* (the stem of a unigram in the independent clause is "be"), *BinPosition0* (this clause appears in the first bin), and *FirstClause* (it is the first clause of the story). An example of an action clause is presented in row 3 with features as follows: *IND-FirstPerson* (it is a first-person independent clause), *POS:IND-VBD* (the verb of the independent clause is past

tense), and *IND-LIWC-Motion* (this unigram in the independent clause belongs to the "Motion" category of LIWC).

4.3 Experiments and Results

We used the extracted features for classification experiments and applied several algorithms, such as Naive Bayes (NB), Confidence Weighted Linear Classifier (CWLC), Maximum Entropy (ME) and a sequential classifier called Conditional Random Fields (CRF) using Weka software (Hall et al., 2005). Using the 40 stories in the training set we calculated the information gain for each feature. For each subset of the highest valued features (in the range of 2^2-2^{12}), we performed a 10-fold cross-validation on the training data and assessed the performance of each classifier to find the right number of features. The feature selection experimental results are shown in Table 4.3.

We report the optimal number of features and the corresponding macro F-score, weighted by the relative frequency of each category for each algorithm. The Naive Bayes (NB) and CRF models perform better with a small subset of the features, while the ME and CWLC algorithms use a much larger subset. Surprisingly the sequential classifier has the lowest F-score and Naive Bayes performs the best. Using the optimal number of features obtained from this search we trained a model for each algorithm using the training dataset and applied these models to the unseen test data and evaluated the performance of each classifier.

We first compute the precision, recall and F-score aggregated over all the clauses in the test set. Table 4.4 summarizes the results for each classifier and label set. The left hand side

Classification Model	CRF	CWLC	Max Entropy	Naive Bayes
# Features	2^7	2^{11}	2^{11}	2^{10}
F1-measure	0.65	0.73	0.73	0.76

Table 4.3: Optimal number of features found for each model and the average F-score obtained using a 10-fold cross-validation on the training data

Model	Overall-P	Overall-R	Overall-F1	Orientation-F1	Evaluation-F1	Action-F1
CRF	0.650	0.665	0.656	0.556	0.742	0.640
CWLC	0.624	0.647	0.632	0.480	0.747	0.609
ME	0.681	0.700	0.688	0.580	0.780	0.670
NB	0.687	0.705	0.689	0.565	0.780	0.687

Table 4.4: Performance of each classifier on the test set when all clauses are aggregated together

of the table shows the macro precision, recall and F-score weighted by the relative frequency of each label. The right hand side of the table shows the F-score of each individual label separately. In this evaluation, Naive Bayes outperforms the other methods and overall, precision and recall are relatively balanced achieving a maximum F-score of 0.689. The CRF does surprisingly well considering its poor performance during the feature selection search. The classifiers perform the poorest on orientation clauses and the best on evaluation clauses.

As mentioned before, the annotation task is highly subjective, requiring interpreting the narrative and the author's intention, which prevents us from obtaining high levels of inter-rater agreement. Because of the noise in the annotations, the standard evaluation metrics may hide information about the ability of the classifiers to learn from our feature set. For example, the best performing classifier (NB) incorrectly labeled 127 clauses out of 430 in the test set. However, 44 of these errors agreed with at least one annotator, but were counted as entirely

Agreement	# Clauses	Prec	Rec	F1	Orientation	Evaluation	Action
1 of 3	15	0.333	0.400	0.339	0.000	0.625	0.333
2 of 3	146	0.597	0.610	0.580	0.472	0.700	0.622
3 of 3	269	0.770	0.773	**0.767**	0.667	0.824	0.746
All	430	0.687	0.705	0.689	0.565	0.780	0.687
Adjusted	430	0.646	0.660	0.643	0.516	0.745	0.623

Table 4.5: Performance measures of Naive Bayes model for different levels of agreement among the annotators

incorrect in the previous evaluations. To address these concerns we also evaluated the performance of the the best performing classifier based on the level of agreement of each instance using two different approaches.

The first approach was inspired by previous work on identifying general vs. specific sentences in news articles (Louis and Nenkova, 2011). They ask five annotators on Amazon Mechanical Turk to label about 600 sentences from WSJ and AP news corpora. They show that the accuracy of the classifier is correlated with the level of agreement between the annotators. For example, their accuracy on WSJ is above 90% for sentences where all five annotators agreed on the label. However, when they include more sentences with lower agreement the accuracy is decreased to 73%. Motivated by this approach we calculate the performance of Naive Bayes classifier for different levels of inter-annotator agreement as shown in Table 4.5. The first three rows show the performance for three different levels of agreement in our dataset. Here 1 means no agreement at all for which there were only 15 clauses in the test set. It is unsurprising that when the annotators could not agree on a label the system performed near chance levels. However, when all three annotators agreed on the gold standard label the F-score improved to

Model	Min-F1	Max-F1	Average-F1
CRF	0.333	0.837	0.609
CWLC	**0.458**	**0.877**	**0.658**
Max Entropy	0.333	0.830	0.649
Naive Bayes	0.333	0.851	0.647

Table 4.6: Summary statistics of the F-score when evaluating stories

0.767. As a comparison, the F-score of the entire test set was 0.689 as shown in the row labeled *All*.

Our second approach is based on the work by Tetreault et al. (2013) on the automated methods for grammatical error detection and correction in writings for ESL (English as a Second Language) tests. They study the problems in annotation and evaluation of ESL error detection systems and present their solutions for improving their performance. They introduce a modification to the standard precision, recall and F-scores that takes into account the level of agreement of each instance, where the values of true-positives and false-negatives are assigned fractional counts based on the proportion of annotators who assigned that label. The final row of Table 4.5 (Adjusted) provides the results using these adjusted values.

We also investigated the performance of the classifiers when evaluating each story separately. Table 4.6 summarizes these results. Each classifier was applied to the clauses of the 10 narratives in the test set and the F-score was computed for each narrative individually. The table shows the minimum, maximum and average F-score over the 10 narratives. The CWLC performed the best on this test and the results show that there is a high variance in performance between stories, with a minimum F-score of 0.458 and a maximum of 0.877. This indicates that some clauses are ambiguous and difficult to label, but also that some *stories* are more difficult

to classify.

4.3.1 Error Analysis

Our results to date indicate that we achieve an overall F-score of 0.689, and that our classifiers are most accurate for the *evaluation* and *action* categories. We also categorized prediction errors for the Naive Bayes classifier based on their sources, finding three major types as shown in Table 4.7 with examples. The most common errors involved clauses that have more than one function in the story, **Multiple Functions**, but may include lexical features that are highly correlated with only one category. For example, the first row of Table 4.7 shows a clause that was labeled as Orientation as it sets the scene for the rest of the story. However, it also describes the narrator's affects and includes features such as **unfortunately**, **could** and **not** that are all strong indicators of evaluation and therefore is predicted as Evaluation by the classifier. The second example of this type shows a clause in which a new character, *Alejandrio*, is introduced and a rising action is described, *trekking to the waterfall*. So both orientation and action are present in this situation. The commonness of such clauses in the corpus suggest that we need a finer grained scheme to capture the multiple levels that such utterances contribute to the story meaning.

Another source of error is when the function of the clause in the narrative is ambiguous like the second row in Table 4.7. While there may be some misleading lexical indicators in these clauses, there were often no strongly correlated words, such as the adjectives and modal verbs in evaluations. The distinction in these cases is that the primary function of the clause within the story is unclear, even to a human reader. Unsurprisingly, most of the examples in this

Error Type	Freq	Gold	Pred	Example
Multiple Functions	61	Ori	Eva	So, **unfortunately** I **couldn't** make the Gamesindustry.biz party tonight.
		Act	Ori	We trekked to a waterfall in the park with the help of Alejandrio a 65 year old Honduran guy who not only walked quicker than us but also carried all the water.
Ambiguity	20	Ori	Eva	I know it is a remarkable haircut because on the way home a handsome young Moroccan man nearly died to tell me how beautiful I was.
Inference	6	Eva	Ori	That's that rabbit food that all of those Northeastern Kerry voters...

Table 4.7: Several common sources of errors with a prototypical example

category had low inter-annotator agreement.

Some of the clauses contain figurative language or complex constructions that require a significant amount of world knowledge and inference to interpret. For example, understanding the clause in row 3 of Table 4.7 requires recognizing the metaphor about rabbit food in order to identify the subjective evaluation the narrator is making. The types of errors described above are not mutually exclusive, for example a clause could be semantically complex and thus have an ambiguous function. The complexity issue, along with ambiguity and multiple functionality of clauses in blog stories are also the reasons why the annotation reliability is not high. In addition, based on this analysis we decided to use our classifier for only separating actions that seemed to be less ambiguous and not further in the rest of the thesis.

4.3.2 Identifying Action vs. Non-Action

One of our initial goals for learning L&W's linear structure is to develop a robust model of narrative events and the relations between them. Based on the L&W's theory, the narrative event causality is described in action clauses. Consequently, we posit that a corpus

Dataset	Overall			Action		
	Precision	Recall	F-1	Precision	Recall	F-1
L&W annotated	0.798	0.802	0.796	0.765	0.625	0.688
Expanded	0.96	0.95	0.95	0.896	0.97	0.932

Table 4.8: Classification results for identifying action vs. non-action

consisting of only actions will allow us to learn stronger contingency relations between events. We tune our features and classifier to develop on a high-precision model for identifying action versus non-action clauses and use it for automatically extracting a corpus of actions from blogs. For this purpose, we first expand our annotated dataset using a heuristic approach. We manually compiled a list of action and non-action verbs from FrameNet and LIWC, as shown in Figure 4.3 and used them to collect more action and non-action clauses as follows:

- **Action Clause:** The root of the clause is an *action* verb and the clause is not negated and the clause is not in *irrealis* mood (e.g. conditional).
- **Non-Action Clause:** The root of the clause is a *non-action* verb.

We collected about 3K action clauses and 6K non-action clauses to expand our corpus and repeated our experiments with the goal of achieving a high precision for the action class. Table 4.8 shows the best results obtained by training a *Logistic Regression* classifier. The first row presents the results of classification using annotated dataset that was described in Section 4.1. The second row shows our results using the expanded dataset which was randomly split into 70% train set and 30% held-out test set. We obtained about 90% precision for the action category with an overall F-1 of 95%.

Action Verbs	ambush, arrive, assail, assault, attack, attend, bend, blink, bomb, bombard, carry, catch, clap, climb, close, compose, crawl, dance, deliver, detect, depart, drive, enter, flee, fly, gaze, glance, glide, goggle, hear, hike, hop, jog, jump, leap, listen, march, move, nod, observe, overhear, print, raid, rehearse, replace, ride, run, sail, say, see, send, shiver, skim, skip, smack, speak, stare, steal, stretch, strike, swim, swing, throw, toddle, travel, visit, walk, watch, witness, write, yawn
Non-Action Verbs	be, dislike, enjoy, feel, hate, like, love, seem, think, want

Figure 4.3: Verb lists used for bootstrapping, collected from FrameNet and LIWC

We will use this classifier to extract action clauses from our Personal Blog Stories corpus for modeling narrative events. This part of our work will be discussed with detailed in Chapter 5. Despite our hypothesis, the results did not indicate any improvements compared to using the entire stories (and not only actions).

4.4 Conclusions

This Chapter presented our work on developing a supervised method for automatically identifying a high-level linear structure of the narratives based on Labov and Waletzky's theory. Our approach included creating a dataset with gold-standard labels based on three categories of narrative clauses: Orientation, Action, and Evaluation. We achieved an overall F-score of 0.689, substantially higher than a majority class baseline with F-score of 0.437. On the data instances with total annotator agreement the F-1 score was increased to 0.767.

The results highlight several properties of the data. Performance is different for results by story rather than over clauses. This indicates that some stories are more difficult to model than others and that ambiguous clauses are not uniformly distributed but are likely to be correlated with particular authors or writing styles.

Our data annotation task requires interpreting the narrative and the author's intention which could be subjective and result in lower levels of inter-annotator agreement. After several refinements of the annotation process, we achieved a kappa score of 0.630. A difficulty of L&W's labeling scheme is the annotation task does not scale well and would be expensive on large amounts of data.

Another limitation of this framework is the ambiguity of the purpose of a sentence in a story as we discussed in Section 4.3.1. Our goal in this work was to derive the L&W's linear model as a foundation to identify finer-grained components and infer the relations between them. However, our analysis of the main sources of classification errors shows that clauses in our corpus can have more than one function in the narrative. For example, a clause in a blog story can describe an event and at the same time express the narrator's affects towards it. This is a major problem with using L&W's model on personal narratives in the blogs dataset, suggesting that it may be too coarse-grained for our purpose.

This theory may be more suitable for narratives that are more strictly structured with a clear flow. For instance, in our primary studies we annotated and experimented with a subset of Aesop's fables. The fables mostly follow a similar structure: a few orientation clauses in the beginning to describe the characters, primary states and location, followed by the clauses reporting the actions and events and ending in a "moral" as the evaluation of the entire story. The linear model defined by L&W could be suitable to apply to narrative data such as the Aesop's fables with more established structures.

For more complex structures however, like what we find in the personal blogs corpus, different approaches with higher level of details are needed to model the narrative elements.

Even with a highly accurate NLP system, the L&W model would not be well-suited for modeling the blog stories. As we discussed in our error analysis and the annotation results in Chapter 3 indicated, even an expert person familiar with the linguistics theories would not be able to easily lay this linear structure over a narrative.

We additionally had a hypothesis that reliably identifying action clauses in the narratives and using them for modeling events can enhance our method for learning contingency relations between events. As we will describe in the next Chapter, our results indicate no improvement when using actions only compared to using the entire narratives. Based on our results and the limitations enumerated above, we will not use L&W's model in the rest of our work on modeling narrative elements. To learn strong contingent event pairs, we propose a different solution and use a weakly-supervised method to generate a topic-specific corpus that could reveal more fine-grained event relations, as we describe in Chapter 5. We then present our methods for identifying human desire expressions in the blog stories in Chapter 6 and describe our experiments on predicting their outcomes.

Chapter 5

Modeling Contingency Relation between Narrative Events

According to the theories of narrative, the story events provide a skeleton for conveying the goals and desires of the characters, as well as the affects, attitudes and themes. The story timeline is then formed by temporal order and causal relations between events. The causality between actions gives the narrative its coherence and contributes to the story timeline as well. In this Chapter, we present our work on identifying narrative events and modeling the relation between them. We focus on specific event relations and use the definition of *Contingency* from Penn Discourse Treebank (PDTB) which has two types: *Cause* relates two events with a cause-effect relation, and *Condition* relates an event with its possible consequences.

We present an unsupervised method for learning such contingency relations between events and directly compare our methods to several other approaches as baselines. Our primary hypothesis is based on the Labov and Waletzky's model that was discussed in the previous

Camping Trip
We packed all our things on the night before Thu (24 Jul) except for frozen food. We brought a lot of things along. **We woke up** early on Thu and JS started packing the frozen marinatinated food inside the small cooler... In the end, we decided the best place to set up the tent was the squarish ground that's located on the right. Prior to setting up our tent, **we placed a tarp on the ground**. In this way, the underneaths of the tent would be kept clean. After that, **we set the tent up**.
Storm
I don't know if I would've been as calm as I was without the radio, as **the hurricane made landfall** in Galveston at 2:10AM on Saturday. As **the wind blew**, branches thudded on the roof or trees snapped, it was helpful to pinpoint the place... **A tree fell** on the garage roof, but it's minor damage compared to what could've happened. We then **started cleaning up**, despite Sugar Land implementing a curfew until 2pm; I didn't see any policemen enforcing this. Luckily my dad has a gas saw (as opposed to electric), so **we helped cut up** three of our neighbors' trees. **I did a lot of raking**, and there's so much debris in the garbage.

Figure 5.1: Excerpts of two stories in the blogs corpus on the topics of *Camping Trip* and *Storm*

Chapter. They suggest that narrative causality is described in the *action* clauses. Therefore, we posit that using the action clauses alone could improve the computational model of events and learning of the contingency relations between them.

As an alternative hypothesis, we propose to use a weakly supervised method for extracting the key events from stories and using them to generate a topic-sorted corpus of personal narratives. We posit that using a topic-specific corpus where the stories share the same theme could reveal more fine-grained and stronger event relations. Examples of personal narratives from two topics, *Camping Trip* and *Storm*, are shown in Figure 5.1.

Previous work has mostly focused on learning of the coarse-grained event relations in newswire genre. Our aim, however, is to learn fine-grained common-sense knowledge about contingent relations between everyday events from personal blog stories. For example, the *Camping Trip* story in Figure 5.1 contains implicit common-sense knowledge about contingent relations between camping-related events, such as *setting up a tent* and *placing a tarp*. The

Storm story contains implicit knowledge about events such as *the hurricane made landfall*, *the wind blew*, *a tree fell*. We use two different datasets to directly compare what is learned from topic-sorted stories as opposed to a general-domain story corpus.

Our results show that the fine-grained knowledge we learn is simply not found in publicly available narrative and event schema collections (Chambers and Jurafsky, 2009; Balasubramanian et al., 2013). In addition, we present new evaluation methods for evaluating contingent event pairs, compared to previous work that mostly uses the Narrative Cloze Test. As a part of our model we identify the indicative contingent event pairs from each topic-specific dataset that can potentially be used as building blocks for generating coherent event chains for a particular theme, however, developing the timeline scaffold of the story is outside of the scope of this thesis.

In Section 5.1 we describe how we learn domain-specific patterns from blog stories and use a weakly-supervised method to identify the indicative events of a specific topic. We use such patterns to expand a hand-labeled corpus of topic-specific personal narratives using a bootstrapping approach as presented in Section 5.2. Our unsupervised method for learning fine-grained contingency relation between events is described in Section 5.3. We describe our experiments and evaluations in Section 5.4. Finally, in Section 5.5 we summarize our work and discuss the conclusions.

5.1 Identifying Indicative Events

We use a weakly-supervised learning algorithm, *AutoSlog-TS*, to identify the typical events that can occur in the stories related to a topic. This method was proposed by Riloff (1996) to automatically generate patterns for information extraction. AutoSlog-TS requires two sets of text data as input; the first set is a collection of documents related to a particular topic, and it needs a second set of random data that are not relevant to the topic of the first set. The algorithm then learns a set of syntactic templates (case frames) that distinguish the linguistic characteristics of the topic-relevant data from the random set.

AutoSlog-TS uses a list of templates for this purpose, some examples of which are shown in the first column of Table 5.1. The example in the first row, *"NP-Prep(NP)"*, means a noun phrase followed by a preposition followed by any noun phrase. The algorithm uses these generic syntactic patterns to learn more specific patterns from its input datasets. The templates are applied to the input documents of both sets in an exhaustive search, to extract lexical patterns matching each template. The second column in Table 5.1 shows such lexico-syntactic patterns extracted from camping trip blog stories. For instance, the first template has matched *"camping in the midst of nature"*, and extracted a domain-specific pattern for camping stories, *" NP-Prep-(NP):POS-CAMPING-IN"*. The second NP that appears in the parenthesis is not a part of the lexico-syntactic template and could be matched to any noun phrase.

Next, the algorithm calculates statistical measures for each pattern based on the labels of the input stories (whether it was from the topic-specific dataset or the random set). These statistics include the frequency and conditional probability indicating how strongly the pattern

Template	Sentence → Lexico-Syntactic Pattern
NP-Prep(NP)	Camping in the midst of nature the midst of nature can significantly bring down stress levels → NP-Prep-(NP):POS-CAMPING-IN
ActVP-Prep(NP)	We set up both tents in record time → ActVP-Prep(NP):POS-SET UP
(subj)-ActVP-Dobj	We have been talking about going camping for a long time now → (subj)-ActVP-Dobj:POS-GOING CAMPING
(subj)-ActVP	When the rain stopped, we packed up camp → (subj)-ActVP:POS-PACKED
(subj)-ActVP	However, while it had a kitchen, we still cooked outside → (subj)-ActVP:POS-COOKED
(subj)-ActVP-Dobj	I went camping a lot as a kid, and those are some of my fondest memories → (subj)-ActVP-Dobj:POS-WENT CAMPING

Table 5.1: Examples of syntactic templates (case frames)

is associated with the topic of the first input set. This algorithm is weakly-supervised as it does not require any annotations of the case frames and the only label needed is one category of the stories on a topic for automatically generating domain-specific patterns.

We manually collected a set of camping trip narratives from blogs corpus and used it as an input dataset along with a random open-domain set to generate domain-specific patterns for this topic. We repeated similar experiment for the *Storm* topic and compiled a set of personal narratives about witnessing a major storm. Some of the learned patterns from each of these two topics are shown in Table 5.2 with their frequency and conditional probability scores. In the first row, *"(subj)-ActVp-Dobj:WENT-CAMPING"* indicates a pattern where the active verb phrase, *went*, is followed by the direct object *camping*. The frequency of this pattern in topic-specific dataset was 21 and its conditional probability is 0.95, indicating that if a document included this pattern, the probability of it being in the topic-specific input set is 95%. The first example for the storm topic is *"(subj)-ActVp-Dobj:LOST-POWER"*. This is interpreted as the active verb phrase, *lost*, followed by its direct object, *power*. The frequency of this pattern in the storm

Topic	Patterns	Frequency	Conditional Prob
Camping Trip	(subj)-ActVp-Dobj:WENT-CAMPING	21	0.95
	NP-Prep-(NP):CAMPING-IN	23	1.0
	NP-Prep-(NP):HIKE-TO	12	0.92
	NP-Prep-(NP):TENT-IN	15	1.0
Storm	(subj)-ActVp-Dobj:LOST-POWER	23	1.0
	(subj)-ActVp:RESTORED	19	0.90
	(subj)-AuxVp-Dobj:HAVE-DAMAGE	17	1.0
	(subj)-ActVp:EVACUATED	9	1.0

Table 5.2: Examples of domain-specific lexico-syntactic patterns

dataset was 23 and its conditional probability is 1.0, specifying that if a story contained this pattern, it is 100% from the storm topic-specific set.

For each topic we generate a list of case frames by using AutoSlog-TS and sort them based on the frequency and conditional probability scores resulting in a list of indicative narrative patterns for each topic. We use this list to develop a classifier and expand our manually labeled topic-specific datasets by bootstrapping as we describe in Section 5.2.

Additionally, we use this list of ranked patterns to identify the *indicative events* of the stories of each topic by choosing the patterns that include a *verb*, since our event representation consists of a verb with its arguments as presented in more details in Section 5.3.1. Table 5.3 shows examples of the indicative events for different topics in our corpus. They are mapped to our event representation, e.g., the pattern *(subj)-ActVP-Dobj:WENT-CAMPING* from Table 5.2 is mapped to *go (dobj:camp)*.

We use the indicative events to identify and evaluate contingent event pairs from each topic-specific corpus as we present in Section 5.4.2, but we never use them for further generating

Topic	Events
Camping Trip	camp(), roast(dobj:marshmallow), hike(), pack(), fish(), go(dobj:camp), grill(), put(dobj:tent , prt:up), build(dobj:fire)
Storm	restore(), lose(dobj:power), rescue(), evacuate(), flood(), damage(), sustain(), survive(), watch(dobj:storm)
Visit a Dentist	infect(), need(dobj:root), yank(), call(dobj:dentist), numb(), drill(), get(dobj:crown), take(dobj:x-ray)
Christmas	open(dobj:present), exchange(dobj:gift), wrap(), sing(), play(), snow(), buy(), decorate(dobj:tree), celebrate()
Snorkeling and Scuba Diving	see(dobj:fish), swim(), snorkel(), sail(), surface(), dive(), dart(), rent(dobj:equipment), enter(dobj:water), see(dobj:turtle)

Table 5.3: Examples of topics in the corpus with some of their indicative events

Last weekend I **went on a "girls only" camping trip** with some friends. I was only just **invited** a week ago, so it was last minute, but I will always **go camping**. This particular trip involved a 7 hour **rafting trip** down the Manistee river, south of Traverse City. What intrigued me was on Wednesday night VanCore **told me** that we would be dressing as pirates during the rafting trip. I did not believe that. It sounded like fun, but I figured I'd **get dressed up** and end up being the only one. But, her sister **confirmed it** without being prompted, so I **packed** my bags, **got out** of work early and **hit the road**. Friday night **we sat** around the campfire **drinking** and **hanging out**. We **put on a ton of fake tattoos** and started **preparing** for the next day. ...It's nice to **wake up** to a campfire. We **got dressed** - bandanas, eye patches, torn up clothing and **grabbed** our dollar store swords and daggers (and our beer grog wench) and **hit the road**... I have never **laughed** so much in my life! - well,I **laughed** until the thunderstorm. Having to **pull over** because you can't even see to steer just isn't that funny. But it IS an adventure...

Figure 5.2: A camping story with highlighted event descriptions

a chain of events and narrative schema. Figure 5.2 shows an example of a camping narrative where the event descriptions are highlighted. Some of these events can be extracted using the indicative patterns, such as *"go camping"* and *"packed"*. However, there are several others that are not the *key events* of a camping trip, such as *"got dressed up"*, *"sat"*, *"hanging out"*, *"wake up"*, and *"laughed"*. This is one of the limitations of our approach, indicating that additional methods and systems are needed for building a complete chain of events from a narrative. Thus we do not tackle the problem of constructing a timeline of the story events in this thesis.

5.2 Bootstrapping a Topic-Sorted Dataset

We use the lexico-syntactic patterns generated by AutoSlog-TS to develop a classifier for bootstrapping a collection of topic-sorted blog stories. First, for each topic we hand-label a small set ($\sim$ 200-300) of stories from the blogs corpus. Then, we use this topic-specific set as an input for AutoSlog-TS along with a random set, to generate the case frames such as the examples presented in Table 5.2.

We use the frequency and probability generated by AutoSlog-TS and apply a threshold for filtering to select a subset of them as indicative patterns strongly associated with the topic. We then developed a classifier using the list of indicative patterns for each topic: each story in the blogs corpus is processed using AutoSlog-TS to extract the case frames and match them with the indicative patterns. If the number of matching indicative patterns is greater than a threshold, the blog story is labeled as a positive instance for that topic.

To find the optimal values for the frequency and probability thresholds, and the matching indicative patterns count, we applied a parameter tuning technique. We divided the hand-labeled data into train and development sets and ran the classifier for several combinations of different values for each of the three parameters. We measured the precision and recall for each iteration and selected the optimal values for the thresholds that resulted in high precision (above 0.9) and average recall (around 0.4). We compromised on a lower recall to achieve a high precision and establish a highly accurate bootstrapping algorithm. Since bootstrapping is performed on a large set of stories, a low recall still results in identifying enough stories per topic.

We used this method to generate a topic-specific corpus of personal stories as de-

scribed in Chapter 3. The manually labeled dataset includes 361 Storm and 299 Camping Trip stories. After one round of bootstrapping the algorithm identified 971 additional Storm and 870 more Camping Trip stories. Examples of stories in this corpus were presented earlier in this Chapter in Figure 5.1. The bootstrapping method is not evaluated separately, however, the results in Section 5.4.1 indicate that using the bootstrapped data considerably improves the accuracy of the contingency model and enhances extracting topic-relevant event knowledge.

5.3 Learning Fine-Grained Contingency Relation between Events

In this section we present our methods for learning fine-grained contingency relations between events from personal narratives. We first describe our event representation and the unsupervised method for learning contingent event pairs. Then we summarize the baseline methods that we use from previous work for comparison. Our evaluations show that the fine-grained knowledge we learn is not found in publicly available narrative and event schema collections.

5.3.1 Event Representation

In previous work different representations have been proposed for the event structure such as single verb and verb with two or more arguments. Verbs are used as a central indication of an event in a narrative. However, other entities related to the verb also play a strong role in conveying the meaning of the events. Pichotta and Mooney (2014) show that the multi-argument representation is richer than the previous ones and is capable of capturing interactions between

multiple events. We use a representation that incorporates the *Particle* of the verb in the event structure in addition to the *Subject* and the *Direct Object* and define an event as a verb with its dependency relations as follows:

Verb Lemma (subj:Subject Lemma, dobj:Direct Object Lemma, prt:Particle)

Table 5.4 shows example sentences describing an event from the Camping topic along with their event structure. The examples show how including the arguments can often change the meaning of an event. In row 1 the *direct object* and *particle* are required to fully understand the event represented by verb *"put"*. Row 2 shows another example where the verb *"have"* cannot implicate what event is happening and the direct object *"oatmeal"* is needed to understand what has occurred in the story.

We parse each sentence and extract every verb lemma with its arguments using Stanford Parser (Manning et al., 2014). For each verb, we extract the *nsubj*, *dobj*, and *prt* dependency relations if they exist, and use their lemma in the event representation. To generalize the event representations, we use the types identified by Stanford's Named Entity Recognizer and map each argument to its named entity type if available, for example, in row 3 of Table 5.4, the *Lost Valley River Campground* is represented by its type LOCATION. We use abstract types for named entities such as PERSON, ORGANIZATION, TIME and DATE. We also represent each pronoun by the abstract type PERSON, such as the example in row 5 of Table 5.4.

5.3.2 Causal Potential Method

We define a *contingent event pair* as a sequence of two events (e_1, e_2) such that e_1 and e_2 are likely to occur together in the given order and e_2 is contingent upon e_1. We apply an

#	**Sentence → Event Representation**
1	but it wasn't at all frustrating *putting up the tent* and setting up the first night → put (dobj:tent, prt:up)
2	The next day *we had oatmeal* for breakfast → have (subj:PERSON, dobj:oatmeal)
3	by the time *we reached the Lost River Valley Campground*, it was already past 1 pm → reach (subj:PERSON, dobj:LOCATION)
4	then *JS set up a shelter* above the picnic table → set (subj:PERSON, dobj:shelter, prt:up)
5	once the rain stopped, *we built a campfire* using the firewoods → build (subj:PERSON, dobj:campfire)

Table 5.4: Event representation examples from Camping Trip stories

unsupervised distributional measure called *Causal Potential* to induce the contingency relation between two events.

Causal Potential (CP) was introduced by Beamer and Girju (2009) as a way to measure the tendency of an event pair to encode a causal relation, where event pairs with high CP have a higher probability of occurring in a causal context. We calculate CP for every pair of adjacent events in each topic-specific dataset. We used a 2-skip bigram model which considers two events to be adjacent if the second event occurs within two or less events after the first one.

We use skip-2 bigram in order to capture the fact that two related events may often be separated by a non-essential event, because of the oral-narrative nature of our data (Rahimtoroghi et al., 2014). In contrast to the verbs that describe an event (e.g., *hike, climb, evacuate, drive*), some verbs describe private states such as as *belong, depend, feel, know*. We filter out clauses that tend to be associated with private states (Wiebe, 1990). A pilot evaluation showed that this improves the results.

Equation 5.1 shows the formula for calculating Causal Potential of a pair consisting of two events: (e_1, e_2). Here, P denotes probability and $P(e_1 \rightarrow e_2)$ is the probability of

e_2 occurring after e_1 in the adjacency window which is equal to 3 due to the skip-2 bigram model. $P(e_2|e_1)$ is the conditional probability of e_2 given that e_1 has been seen in the adjacency window.

$$CP(e_1, e_2) = log\frac{P(e_2|e_1)}{P(e_2)} + log\frac{P(e_1 \rightarrow e_2)}{P(e_2 \rightarrow e_1)} \tag{5.1}$$

To calculate CP, we need to compute event counts from the corpus and thus we need to define when two events are considered equal. The simplest approach is to define two events to be equal when their verb and arguments exactly match. However, with a closer look at the data this approach does not seem adequate. For example, consider the following events:

go (subj:PERSON, dobj:camp)
go (subj:family, dobj:camp)
go (dobj:camp)

They encode the same action although their representations do not exactly match and differ in the subject. Our intuition is that when we count the number of events represented as *go (subj:PERSON, dobj:camp)* we should also include the count of *go (dobj:camp)*. To be able to generalize over the event structure and take into account these nuances, we consider two events to be equal if they have the same verb lemma and share at least one argument other than the subject.

5.3.3 Baseline Methods

Our previous work on modeling contingency relations in film scripts data compared Causal Potential to methods used in previous work: Bigram event models (Manshadi et al.,

2008) and Pointwise Mutual Information (PMI) (Chambers and Jurafsky, 2008) and the evaluations showed that CP obtains better results (Hu et al., 2013). Therefore, we use CP for inducing contingency relation between events in the blogs and compare it to three other models as baselines:

Event-Unigram. This method will produce a distribution of normalized frequencies for events.

Event-Bigram. We calculate the bigram probability of every pair of adjacent events using skip-2 bigram model using the Maximum Likelihood Estimation (MLE) from our datasets:

$$P(e_2|e_1) = \frac{Count(e_1, e_2)}{Count(e_1)} \tag{5.2}$$

Event-SCP. We use the Symmetric Conditional Probability between event tuples (Rel-grams) used by Balasubramanian et al. (2013) as another baseline method. The Rel-gram model is the most relevant previous work to our method and outperforms the previous state of the art on generating narrative event schema. This metric combines bigram probability considering both directions:

$$SCP(e_1, e_2) = P(e_2|e_1) \times P(e_1|e_2) \tag{5.3}$$

Like Event-Bigram, we used MLE for estimating Event-SCP from the corpus.

5.4 Experiments and Evaluations

We propose two evaluation methods for our work and conduct different sets of experiments to evaluate different aspects of our model. First, we develop an automatic two-choice test, modeled after the COPA task (Roemmele et al., 2011), and run it on a held-out test set

Topic	Dataset	# Docs
Camping Trip	Hand-labeled held-out test	107
	Hand-labeled train (Train-HL)	192
	Train-HL + Bootstrap (Train-HL-BS)	1,062
Storm	Hand-labeled held-out test	98
	Hand-labeled train (Train-HL)	263
	Train-HL + Bootstrap (Train-HL-BS)	1,234

Table 5.5: Number of stories in the train and test sets of the topic-specific dataset

to evaluate the event pair collections that we have extracted. We created a *General-Domain* dataset for comparison to Topic-Specific results. It is a random subset of the Personal Blog Stories corpus consisting of 4,200 stories not selected for any specific topic.

Second, we create an experiment on Amazon Mechanical Turk (AMT) to evaluate the top N topic-specific event pairs with respect to their contingency relation and topic-relevance. Finally, we compare the content of our topic-specific event pairs to current state of the art event collections to show that the fine-grained knowledge we learned about everyday events does not exist in previous work focused on the news genre.

5.4.1 Automatic Two-Choice Test

For evaluating our contingent event pair collections we have automatically generated a set of two-choice questions along with the answers, modeled after the COPA task (Roemmele et al., 2011). We produced questions from held-out test sets for each dataset. Each question consists of one event and two choices. The *question event* is one that occurs in the test data. One of the choices is an event adjacent to the question event in the document. The other choice

is an event randomly selected from the list of all events occurring in the test set. The following is an example of a question from the Camping Trip test set:

Question event: arrange (dobj:outdoor)
Choice 1: help (dobj:trip)
Choice 2: call (subj:PERSON)

In this example, *arrange (dobj:outdoor)* is followed by the event *help (dobj:trip)* in a document from the test set and *call (subj:PERSON)* was randomly generated. The model is supposed to predict which of the two choices is more likely to have a contingency relation with the event in the question. We argue that a strong contingency model should be able to choose the correct answer (the one that is adjacent to the question event) and the accuracy achieved on the test questions is an indication of the model's robustness.

For the General-Domain dataset, we split the data into train (4,000 stories) and held-out test (200 stories) sets. For each topic-specific set, we divided the hand-labeled data into a train (Train-HL) and held-out test, and created a second train set consisting of Train-HL and the data collected by bootstrapping (Train-HL-BS) as shown in Table 5.5. We automatically created a question for every event occurring in the test data which resulted in 3,123 questions for General-Domain data, 2,058 for the Camping and 2,533 questions for the Storm topic.

5.4.1.1 Experiment on General-Domain Dataset

We applied the baseline methods and Causal Potential model on the train set to learn contingent event pairs and tested the pair collections on the questions generated from held-out test set. We extracted about 418K contingent event pairs from General-Domain train set and used our automatic test approach to evaluate these event pair collections. The results are shown

Model	Accuracy
Event-Unigram	0.478
Event-Bigram	0.481
Event-SCP (Rel-gram)	0.477
Causal Potential	0.510

Table 5.6: Automatic two-choice test results for General-Domain dataset

in Table 5.6. Figure 5.3 shows some examples of event pairs with high CP scores extracted from general-Domain set.

Our primary hypothesis based on the Labov and Waletzky's theory was that a corpus consisting of only actions will allow us to learn stronger contingency relations between events, as narrative event causality is described in action clauses. As we described in Chapter 4, we developed a high performing classifier for identifying actions with an overall F-1 of 95% and about 90% precision for the action category.

We used this classifier to extract action clauses from our general-domain train set and created a dataset and applied our methods for learning contingency relations between events on this dataset. However, the results did not indicate any improvements compared to using the entire stories (and not only actions) presented in Table 5.6. Consequently, we propose a different approach for addressing this problem and generated a topic-specific corpus of blogs for learning fine-grained contingent event pairs as we describe next.

5.4.1.2 Experiment on Topic-Specific Datasets

For each topic-sorted dataset, we applied the Causal Potential model on the train sets to learn contingent event pairs and extracted about 437K event pairs from Storm Train-HL-BS

1	go (nsubj:PERSON) → go (dobj:trail , prt:down)
2	find (nsubj:PERSON , dobj:fellow) → go (prt:back)
3	see (nsubj:PERSON , dobj:gun) → see (dobj:police)
4	go (nsubj:PERSON) → go (nsubj:PERSON , dobj:rafting)
5	come (nsubj:PERSON) → go (nsubj:PERSON)
6	go (prt:out) → find (nsubj:PERSON , dobj:sconce)
7	go (nsubj:PERSON) → see (dobj:window, prt:out)
8	go (nsubj:PERSON) → walk (dobj:bit , prt:down)

Figure 5.3: Examples of event pairs with high CP scores extracted from General-Domain stories

and 630K pairs from Camping Trip Train-HL-BS set. We also applied the baseline methods on each dataset and used our automatic test approach to evaluate these event pair collections on the questions generated from held-out test set. The evaluation results are shown in Table 5.7.

The Causal Potential model trained on Train-HL-BS dataset achieved accuracy of 0.685 on Camping Trip and 0.887 on Storm topic which is stronger than all the baselines. Our experiments indicate that having more training data collected by bootstrapping improves the accuracy of the model in predicting contingency relation between events. Additionally, the Causal Potential results on Topic-Specific dataset is stronger than General-Domain narratives indicating that using a topic-sorted dataset improves learning causal knowledge about events. In the following section we extract topic-indicative contingent event pairs and show that Topic-Specific data enables learning of finer-grained event knowledge that pertain to a particular theme.

5.4.2 Topic-Indicative Contingent Event Pairs

We identify contingent event pairs that are highly indicative of a particular topic. We hypothesize that these event pairs serve as building blocks of coherent event chains and narrative schema since they encode contingency relation and correspond to a specific theme. We evaluate

Topic	Model	Train Dataset	Accuracy
Camping Trip	Event-Unigram	Train-HL-BS	0.507
	Event-Bigram	Train-HL-BS	0.510
	Event-SCP	Train-HL-BS	0.508
	Causal Potential	Train-HL	0.631
	Causal Potential	**Train-HL-BS**	**0.685**
Storm	Event-Unigram	Train-HL-BS	0.510
	Event-Bigram	Train-HL-BS	0.523
	Event-SCP	Train-HL-BS	0.516
	Causal Potential	Train-HL	0.711
	Causal Potential	**Train-HL-BS**	**0.887**

Table 5.7: Automatic two-choice test results for Topic-Specific dataset

the pairs on Amazon Mechanical Turk (AMT).

To identify event sequences that have a strong correlation to a topic (topic-indicative pairs) we applied two filtering methods. First, we selected the frequent pairs for each topic and removed the ones that occur less than 5 times in the corpus. Second, we used the indicative event-patterns for each topic and extracted the pairs that at least included one of these patterns. Indicative event-patterns are automatically generated during the bootstrapping using AutoSlog-TS and mapped to their corresponding event representation as described in Section 5.1. Then we used the Causal Potential scores from our contingency model for ranking the topic-indicative event pairs to identify the highly contingent ones. We sorted the pairs based on the Causal Potential score and evaluated the top N pairs in this list.

5.4.2.1 Evaluation on Mechanical Turk

We evaluate the indicative contingent event pairs using human judgment on Amazon Mechanical Turk (AMT). Narrative schema consists of chains of events that are related in a coherent way and correspond to a common theme. Consequently, we evaluate the extracted pairs based on two main criteria:

- **Contingency:** Two events in the pair are likely to occur together in the given order and the second event is contingent upon the first one.

- **Topic Relevance:** Both events strongly correspond to the specified topic.

We have designed one task to assess both criteria since if an event pair is not contingent, it cannot be used in narrative schema for not satisfying the required coherence (even if it is topic-relevant). We asked the AMT annotators to rate each pair on a scale of 0-3 as follows:

0: The events are not contingent.
1: The events are contingent but not relevant to the specified topic.
2: The events are contingent and somewhat relevant to the specified topic.
3: The events are contingent and strongly relevant to the specified topic.

To ensure that the Amazon Mechanical Turk annotations are reliable, we designed a *Qualification Type* which requires the workers to pass a test before they can annotate our pairs. If the workers score 70% or more on the test they will qualify to do the main task. For each topic we created a Qualification test consisting of 10 event pairs from that topic that were annotated by two experts. To make the events more readable for the annotators we used the following representation:

Subject - Verb Particle - Direct Object

For example, *hike(subj:person, dobj:trail, prt:up)* is mapped to *person - hike up - trail*. For each topic we evaluated top $N = 100$ event pairs and assigned 5 workers to rate each one. We generated a gold standard label for each pair by averaging over the scores assigned by the annotators and interpreted the average as follows:

Label $>$2: Contingent & strongly topic-relevant.

Label = 2: Contingent & somewhat topic-relevant.

1 $\leq$ Label $<$ 2: Contingent & not topic-relevant.

Label $<$ 1: Not contingent.

To assess the inter-annotator reliability we calculated kappa between each worker and the majority of the labels assigned to each pair. The average kappa was 0.73 which indicates substantial agreement. The results in Table 5.8 show that 52% of the Camping Trip and 53% of the Storm pairs were labeled as contingent and topic-relevant by the annotators. The results also indicate that our model is capable of identifying event pairs with strong contingency relations: 82% of the Camping Trip pairs and 77% of the Storm pairs were marked as contingent by the workers. Examples of the strongest and weakest pairs evaluated on Mechanical Turk are shown in Table 5.9. By comparison to Figure 5.3, we can see that we can learn finer-grained type of events knowledge from topic-specific stories as compared to general-domain corpus.

Label	Camping	Storm
Contingent & Strongly Relevant	44 %	33 %
Contingent & Somewhat Relevant	8 %	20 %
Contingent & Not Relevant	30 %	24 %
Total Contingent	82 %	77 %

Table 5.8: Results of evaluating indicative contingent event pairs on AMT

Topic	Label $>$ 2 : Contingent & Strongly Topic-Relevant	Label $<$ 1 : Not Contingent
Camping Trip	person - pack up → person - go - home person-wake up → person-pack up-backpack person - head → hike up climb → person - find - rock person - pack up - car → head out	person - pick up - cup → person - swim pack up - tent → check out - video person - play → person - pick up - sax pack up - material → switch off - projector person - pick up - photo → person - swim
Storm	wind - blow - transformer → power - go out tree - fall - eave → crush Ike - blow → knock down - limb air - push - person → person - fall out hit - location → evacuate - person	restore - community → hurricane - bend boil → tree - fall - driveway clean up - person → people - come out blow - sign → person - sit person - rock - way → bottle - fall

Table 5.9: Examples of event pairs evaluated on AMT

5.4.3 Comparison to Rel-gram Tuple Collections

We chose Rel-gram tuples (Balasubramanian et al., 2013) for comparison since it is the most relevant previous work to us: they generate pairs of relational tuples of events, called *Rel-grams* using co-occurrence statistics based on Symmetric Conditional Probability described in Section 5.3.3. Additionally, the Rel-grams are publicly available through an online search interface[1] and their evaluations show that their method outperforms the previous state of the art

[1] http://relgrams.cs.washington.edu:10000/relgrams

Label	Rel-gram Tuples
Contingent & Strongly Relevant	7 %
Contingent & Somewhat Relevant	0 %
Contingent & Not Relevant	35 %
Total Contingent	42 %

Table 5.10: Evaluation of Rel-gram tuples on AMT

on generating narrative event schema.

However, their work is focused on news articles and does not consider the causal relation between events for inducing event schema. We compare the content of what we learned from our topic-specific corpus to the Rel-gram tuples to show that the fine-grained type of knowledge that we learn is not found in their events collection. We also applied the co-occurrence statistics that they used on our data as a baseline (Event-SCP) for comparison to our method (the results were presented in Section 5.4.1).

In this experiment we compare the event pairs extracted from our Camping Trip topic to the Rel-gram tuples. The Rel-gram tuples are not sorted by topic. To find tuples relevant to Camping Trip, we used our top 10 indicative events and extracted all the Rel-gram tuples that included at least one event corresponding to one of the Camping Trip indicative events. For example, for *go(dobj:camp)*, we pulled out all the tuples that included this event from the Rel-grams collection. The indicative events for each topic were automatically generated during the bootstrapping using AutoSlog-TS (Section 5.1).

Then we applied the same sorting and filtering methods presented in the Rel-grams work and removed any tuple with frequency less than 25 and sorted the rest by the total symmet-

rical conditional probability. These numbers are publicly available as a part of the Rel-grams collection. We evaluated the top $N = 100$ tuples of this list using the Mechanical Turk task described in Section 5.4.2. The evaluation results presented in Table 5.10 show that 42% of the Rel-gram pairs were labeled as contingent by the annotators and only 7% were both contingent and topic-relevant. We argue that this is mainly due to the limitations of the newswire data which does not contain the fine-grained everyday events that we have extracted from our personal blogs corpus.

5.5 Summary

We presented our work aiming to learn fine-grained common-sense knowledge about contingent relations between everyday events from personal blog stories. This type of knowledge is useful for information extraction, question answering, and text summarization. We used a weakly-supervised algorithm to identify the indicative events from domain-specific personal narratives and applied a bootstrapping approach using these event patterns to create topic-sorted sets of stories. We developed a method for learning contingency relations between events and evaluated our methods on a set of general-domain narratives as well as two topic-specific datasets. Our evaluations indicate that a method that works well on the news genre does not generate coherent results on personal stories (comparison of Event-SCP baseline with Causal Potential).

We modeled the contingency (causal and conditional) relation between the events from each dataset using Causal Potential and evaluated on the questions automatically generated

from a held-out test set. The results show significant improvement over the Event-Unigram, Event-Bigram, and Event-SCP (Rel-grams method) baselines on Topic-Specific stories: 25% improvement of accuracy on Camping Trip and 41% on Storm topic compared to Bigram model. One of the future directions of our work is to explore existing topic-modeling algorithms to create a broader set of topic-sorted corpora for learning contingent event knowledge.

Our comparison to Rel-gram tuples shows that our model outperforms their method when evaluated in terms of the contingency relation between events and relevance to the topic. Our experiments show that most of the fine-grained contingency relations we learn from narrative events are not found in existing narrative and event schema collections induced from the newswire datasets (Rel-grams).

We also extracted indicative contingent event pairs from each topic and evaluated them on Mechanical Turk. The evaluations show that 82% of the relations between events that we learn from topic-sorted stories are judged as contingent. Another future line of work could be using the contingent event pairs as building blocks for generating coherent event chains and narrative schema. Our results indicate that using topic specific datasets enables learning fine-grained knowledge about events and results in significant improvement over the baselines.

One of the limitations of our approach is that we only generate the indicative events of the stories and aim to model the relation between pairs of events. We do not construct a timeline of the narrative events and additional methods are still needed for building a complete chain of events. However, as we have discussed previously, the story events provide a skeleton for conveying the other parts of the narrative structure, such as goals, emotions and affective states. Understanding a narrative also involves modeling the protagonists' goals and desires as

a way of explaining their motivations and their actions. Therefore, we conclude our work on modeling the narrative events in this Chapter and aim to study the desires of the protagonist and their outcome in Chapter 6.

Chapter 6

Modeling Goals and Desires in Narratives

In this Chapter we present the final part of our work in modeling the narrative structure. We propose methods for recognizing the expression of the protagonist's goals and desires in narratives and tracking their corresponding outcomes, as a sub-problem of modeling affects. Our work is focused on the explicit verbal patterns of desire expressions in personal blog stories.

An example of a desire expression in our corpus with its narrative context is shown in Figure 6.1. Our approach builds on seminal work on a computational model of Lehnert's plot units, that applied modern NLP tools to tracking narrative affect states in Aesop's Fables (Goyal et al., 2010; Lehnert, 1981; Goyal and Riloff, 2013). Our framing of the problem is also inspired by recent work that identified three forms of desire expressions in short narratives from MCTest and SimpleWiki and developed models to predict whether the desires are fulfilled or not (Chaturvedi et al., 2016).

We introduce a new corpus **DesireDB** of ∼3,500 first-person informal narratives with annotations for desires and fulfillment status, as described in Chapter 3. DesireDB aims to pro-

Prior-Context: (1) I ran the Nike+ human Race 10K new York in under 57 minutes! (2) Then at the all-American rejects concert, I somehow ended up right next to this really cute guy and he seemed interested in me. (3) Was I imagining things? He was really nice; (4) I dropped something and it was dark, he bent with his cell phone light to help me look for it. (5) We spoke a little, but it was loud and not suited for conversation there.
Desire-Expression-Sentence: I **[had hoped to]** ask him to join me for a drink or something after the show (if my courage would allow such a thing) but he left before the end and I didn't see him after that.
Post-Context: (1) Maybe I'll try missed connections lol. (2) I didn't want to tell him I think he's cute or make any gay references during the show because if I was wrong that would make standing there the whole rest of the concert too awkward... (3) Afterward, I wandered through the city making stops at several bars and clubs, met some new people, some old people (4) As in people I knew - I actually didn't met any old people, unless you count the tourist family whose dad asked me about my t-shirt. (5) And when I thought the night was over (and the doorman of the club did insist it was over) I met this great guy going into the subway.

Figure 6.1: A desire expression with its surrounding context extracted from a personal blog story

vide a testbed for modeling desire and goals in personal narrative and predicting their fulfillment status. An instance of this corpus is shown in Figure 6.2, where the first part is the extracted data from corpus including the desire expression with prior and post context, and the second part is the gold-standard annotations. The narrative and sentence structure in DesireDB is more complex than either MCTest or SimpleWiki (Richardson et al., 2013; Coster and Kauchak, 2011). The size of the vocabulary in DesireDB is more than 32K, while this number is about 1.8K and 12K for MCTest and SimpleWiki respectively.

We present our methods and features in Section 6.1, with our experiments and results of comparison to previous work. Our experiments and annotation analysis indicated that identifying the unfulfillment of a goal is a harder task compared to the fulfilled ones. Consequently, we perform further analysis using a corpus of thwarted goal expressions, *ThwartDB* (as described in Chapter 3) and present our studies in Section 6.2. Finally, we summarize our work and present our conclusions in Section 6.3.

Data-Instance:
Prior-Context: ConnectiCon!!! Ya baby, we did go this year as planned! Though this year we weren't in the artist colony, so I didn't see much point in posting about it before hand.
Desire-Expression-Sentence: This year we [wanted to] be part of the main crowd.
Post-Context: We wanted to get in on all the events and panels that you cant attend when watching over a table. And this year we wanted to cosplay! My hubby and I decided to dress up like aperture Science test subjects from the PC game portal. It was a good and original choice, as we both ended up being the only portal related people in the con (unless there were others who came late in the evening we didn't see) It was loads of fun and we got a surprising amount of attention.

Annotations:
Fulfillment-Label: Fulfilled
Fulfillment-Agreement-Score: 3
Evidence: Though this year we weren't in the artist colony. We wanted to get in on all the events and panels that you cant attend when watching over a table.
Evidence-Overlap-Score: 3

Figure 6.2: Example of data in DesireDB

6.1 Modeling Desire Fulfillment

The purpose of our work is to predict the fulfillment status (Fulfillment-Label in Figure 6.2) of the desire expressions in DesireDB corpus, given the prior and post context. We conduct a range of experiments on predicting the status of the desires and goals, using different features and models, including LSTM architectures that can encode the sequential structure of the narratives. We first describe our features and models. Then, we present our feature analysis study to examine their importance in modeling fulfillment. Finally, we provide results of direct comparison to previous work on the existing corpora.

6.1.1 Features Description

In our original informal examination of the DesireDB development data, we noticed several ways that a writer can signal (lack of) fulfillment of a desire like "I hoped to pick up a dictionary". First, they may mention an outcome that entails ("The book I bought was...")

or strongly implies fulfillment ("I went back home happily."). However, we noticed that in many cases of fulfillment, the 'marker' was simply the absence of any mention that things went wrong. For lack of fulfillment, while we found cases where writers explicitly state that their desire wasn't met, we noted many instances where evidence came from mentioning that an enabling condition for fulfillment wasn't met ("The bookstore was closed.").

True machine understanding of these kinds of narrative structures requires robust models of the complex interplay of semantics (including negation) as well as world knowledge about the scripts for tasks like buying books, including what count as enabling conditions and entailers for fulfillment. For our experiments we considered reasonable proxies for the conditions mentioned above using existing resources (note that we also tested LSTM models described below, which may implicitly learn such relationships with sufficient data).

We define four different sets of features to capture different aspects of the data and our task. One set (**Desire-Features**) indexes properties of the desire expression (e.g., the desire verb) as well as overlap between the desired object/event and the surrounding context. The remaining features attempt to find general markers for success or failure. We design a set of (**Discourse-Features**) to look for overt discourse relation markers that signal violation of expectation (e.g., 'but', 'however') or its opposite (e.g., 'so') and could be useful in predicting the outcome of a desire.

The polarity of the desire expression and its narrative context could be another signal for fulfillment or unfulfillment. Thus, the (**Connotation-Features** are defined to use the Connotation Lexicon (Feng et al., 2013) and model whether the context provides a positive or negative event. All of these features are inspired by Chaturvedi et al. (2016). Finally, motivated by the

Sentiment: Negative Prior-Context(4): "I had been working for hours on boring paperwork and financial stuff, and I was really crabby."
Sentiment: Negative Prior-Context(5): I decided it was time to take a break and thought, should I read a magazine or watch best Week Ever?
Sentiment: Negative Desire-Expression-Sentence: But I realized that what I really **[wanted to]** do was go for a run!
Sentiment: Positive Post-Context(1): That was pretty amazing, to transition mentally from 'having to' to 'wanting to' run.
Sentiment: Positive Post-Context(2): So I did a quick, fun 2.75 miles.

Figure 6.3: Example of sentiment features, where prior context is negative while the post context is positive, implying fulfillment of the desire

AESOP modeling of affect states for identifying plot units (Goyal and Riloff, 2013), we define a set of **Sentiment-Flow-Features**. Plot units model suggest that the failure of a goal is expressed through a transition from a positive state to a negative one. To capture this, we extract features indicating whether there has been a change in sentiment in the surrounding context which might be the mention of a thwarted effort or a hard won victory, as shown in Figure 6.3.

In addition to a BOW (Bag of Words) baseline, we extracted the four types of features mentioned above. For features that examine the context around the desire expression, our experiments used the prior-context, the post-context, or both, as discussed below; context features are computed per sentence i of the context. We also tested various ablations of these features. We describe the full set of features in more detail below.

1. **Desire-Features.** From a desire expression of the form 'X Ved S', we extract the lexical feature *Desire-Verb*, the lemma for V. We also extract a list of *focal words*, the content words in embedded sentence S. In Figure 6.3, these are 'do', 'go', and 'run' for the desire

expression *I really [wanted to] do was go for a run!*. We count how many times each word, its synonyms, or its antonyms in WordNet (Fellbaum, 1998) are in the context, respectively. Similarly, we identify whether subject X is mentioned in the context and if X is first person (e.g. 'I', 'we').

2. **Discourse-Features.** This class of features count how many of two classes of discourse relation markers (*Violated-Expectation* vs. *Meeting-Expectation*) occur in the context. For the classes, we manually coded all overt discourse relation markers in the Penn Discourse Treebank three ways(violation, meeting, or neutral), leading to 15 meeting markers ('accordingly', 'so', 'ultimately', 'finally') and 31 violating ('although', 'rather', 'yet', 'but'). In addition, we also tracked the presence of the most frequent of these ('so' and 'but', respectively) in the desire sentence.

3. **Connotation-Features.** Beyond the use of WordNet expansion for focal words, we also used the Connotation Lexicon (Feng et al., 2013), a lexical resource marking very general connotation polarities (positive or negative) of words (as opposed to more specific sentiment lexicons). We count the number of words in the context that have the same connotation polarity for each focal word (Connotation-Agree-Feature) and the ones that have the opposite polarity (Connotation-Disagree-Feature).

4. **Sentiment-Flow-Features.** To model affect states, we compute a sentiment score for the desire expression sentence as well as each sentence in the context. Then for each sentence of the context, we define the *Sentiment-Agree* and *Sentiment-Disagree* features to indicate whether that sentence and the desire expression sentence have the same sentiment

Fulfilled	Unfulfilled	Unknown	None	Total
1,366	953	380	70	2,780

Table 6.1: Simple-DesireDB dataset

Method	Features	Ful-P	Ful-R	Ful-F1	Unf-P	Unf-R	Unf-F1	Precision	Recall	F1
Skip-Thought	BOW	0.75	0.70	0.72	0.54	0.61	0.57	0.65	0.65	0.65
	ALL	**0.80**	**0.71**	**0.75**	**0.59**	**0.70**	**0.64**	**0.70**	**0.70**	**0.70**
CNN-RNN	BOW	0.75	0.73	0.74	0.57	0.60	0.58	0.66	0.66	0.66
	ALL	0.75	0.79	0.77	0.61	0.56	0.59	0.68	0.68	0.68

Table 6.2: Results of LSTM models on Simple-DesireDB

polarity (see Figure 6.3). We use the Stanford Sentiment Analyzer (Socher et al., 2013) for extracting these features.

6.1.2 Classification Models and Experiments

Our features are motivated by narrative characteristics but do not directly capture the sequential structure of the narratives. We thus apply neural network models suitable for sequence learning, in order to directly encode the order of the sentences in the story and distinguish between prior and post context. We use two different architectures of LSTM (Long Short-Term Memory) (Hochreiter and Schmidhuber, 1997) models to generate sentence embeddings and then apply a three-layer RNN (Recurrent Neural Network) for classification. We used Keras (Chollet, 2015) as a deep learning toolkit for implementing our experiments.

Skip-Thoughts. This is a sequential model that uses pre-trained skip-thoughts model (Kiros et al., 2015) as the embedding of sentences. It first concatenates features, if any, with embeddings, and then uses LSTM to generate a single representation for the context sequence, which

is the output of the last unit. That single representation is then concatenated with embedding-feature concatenation of desire sentence and is fed into a multi-layer network to yield a single binary output.

CNN-RNN. The only difference between the CNN-RNN model and Skip-Thought is that it uses the 1-dimensional convolution with max-over-time pooling introduced in (Kim, 2014) to generate the sentence embedding from word embedding, instead of using skip-thoughts. We use Google News Vectors (Mikolov et al., 2013) for the word embedding with different sizes from 1 to 7 for the kernel.

For our experiments, we first constructed a subset of DesireDB that we will call **Simple-DesireDB**, in order to be able to compare more directly to the models and data used in previous work. Previous work (Chaturvedi et al., 2016) used three verb phrases to identify desire expressions (*wanted to*, *hoped to*, and *wished to*), so we selected a portion of our corpus including these patterns along with two other expressions (*couldn't wait to* and *decided to*) to have sufficient data for experiments. Table 6.1 shows the distribution of labels in this subset. For classification experiments we use data labeled as *Fulfilled* and *Unfulfilled*, thus the majority class accuracy is 59%. We split the data into Train (1,656), Dev (327), and Test (336) sets for the experiments.

Results of our two LSTM models for Fulfilled (Ful) and Unfulfilled (Unf) classes and the overall classification task (P:precision, R:recall) on Simple-DesireDB are presented in Table 6.2. ALL feature set includes all the features described in Section 6.1.1 (excluding BOW). These features aim to capture the structure of the narrative discourse and semantics of the stories and our classification results show that they can considerably improve the model,

compared to the BOW baseline (F1 improved from 0.65 to 0.70 for Skip-Thought). For instance, the sentiment and connotation features could capture the affect state transitions in the story as shown in Figure 6.3. In this example, the affectual state of the character is negative in the prior context (*"I had been working for hours on boring paperwork..."* and *"I decided it was time to take a break..."*). After the desire expression, it transitions into positive state in the post context (*"That was pretty amazing..."*), signaling the satisfaction of the desire. Another example is the discourse features that model the discourse relations between different parts of context and the desire expression, indicating violation or meeting of the goals.

We also conducted four sets of experiments to study the importance of prior, post and the entire context in predicting fulfillment status, using our best model. The results of Skip-Thought using different contextual representations are in Table 6.3 with ALL features. The results indicate that adding features from prior context **alone** improves the results and thus the description of a goal's outcome could precede the expression of the goal in the narrative. Our best results are obtained by including the post context as well and using features from the desire sentence with the entire surrounding context. We conclude that the evaluations of the goals can be indicated both before and after the explicit goal expression in the narrative.

We then experimented with our best model on all of DesireDB to see if the features and models can identify the goal fulfillment on a wider range of verbal expressions. We trained Naive Bayes, SVM and Logistic Regression (LR) classifiers as baselines, with the best results on the Dev set achieved by Logistic Regression. Table 6.4 shows the results of Skip-Thought and LR on DesireDB for different features on the test set. We performed feature ablation study using Logistic Regression classifier on the Dev set which will be discussed in Section 6.1.3 in

Data	Ful-P	Ful-R	Ful-F1	Unf-P	Unf-R	Unf-F1	Precision	Recall	F1
Desire	0.74	0.75	0.75	0.57	0.56	0.57	0.66	0.66	0.66
Desire+Prior	0.78	0.73	0.75	0.58	0.65	0.61	0.68	0.69	0.68
Desire+Post	0.76	0.70	0.73	0.55	0.62	0.59	0.66	0.66	0.66
Desire+Context	**0.80**	**0.71**	**0.75**	**0.59**	**0.70**	**0.64**	**0.70**	**0.70**	**0.70**

Table 6.3: Results of Skip-Thought using different parts of data, with ALL features on Simple-DesireDB

Method	Features	Ful-P	Ful-R	Ful-F1	Unf-P	Unf-R	Unf-F1	Precision	Recall	F1
Skip-Thought	BOW	0.78	0.78	0.78	0.57	0.56	0.57	0.67	0.67	0.67
	ALL	0.78	0.79	0.79	0.58	0.56	0.57	0.68	0.68	0.68
	Discourse	**0.80**	**0.79**	**0.80**	**0.60**	**0.60**	**0.60**	**0.70**	**0.70**	**0.70**
Logistic Regression	BOW	0.69	0.65	0.67	0.53	0.57	0.55	0.61	0.61	0.61
	All	0.79	0.70	0.74	0.52	0.64	0.58	0.66	0.67	0.66
	Discourse	0.75	0.84	0.80	0.60	0.45	0.52	0.67	0.65	0.66

Table 6.4: Results of best LSTM model with different feature sets, compared to Logistic Regression on DesireDB

details. The results indicated that Discourse features are better predictors of fulfillment status that the other three sets. Therefore, we present results using only Discourse features in addition to BOW and ALL for the Skip-Thought model to examine their effectiveness in a sequential classifier as well.

The discourse features achieve a higher performance compared to using ALL, suggesting that exploiting the discourse structure in the stories could improve the performance of models for understanding the narrative. In addition, comparing the results of Skip-Thought using ALL features on DesireDB to Simple-DesireDB (Tables 6.4 and 6.2) indicates that when we expand the data to include more desire expression patterns the problem becomes more difficult

Features	Precision	Recall	F1
ALL	0.64	0.64	0.64
Discourse	0.66	0.64	0.65
But-Present	**0.72**	**0.64**	**0.68**
ALL w/o But-Present	0.58	0.58	0.58

Table 6.5: Results of Logistic Regression with different feature sets on Simple-DesireDB

and the performance drops. We believe more complicated language discourse structure models are needed for better narrative comprehension.

All of the results indicate that similar features and methods achieve better results for the *Fulfilled* class as compared to *Unfulfilled*. We believe the reason is that identifying unfulfillment of a desire or goal is a more difficult task, as discussed in the annotation description in Chapter 3. To further our analysis on the annotation disagreements, we examined the cases where only two annotators agreed on the assigned label. From the expressions labeled **Fulfilled** by two annotators, 64% were labeled *Unknown from the context* by the disagreeing annotator, and only 36% were labeled *Unfulfilled*. However, these numbers for the **Unfulfilled** class are respectively 49% and 51%, indicating a stronger disagreement between annotators when labeling Unfulfilled expressions.

6.1.3 Feature Selection Experiments

We used the InfoGain measure to rank features based on their importance in modeling desire fulfillment. The top 5 features are: But-Present, Post-Context-Connotation-Disagree, Post-Context-Violated-Expectation, Desire-Verb, Subject-First-Person. We also tested different feature sets separately. We describe our experiment results below.

The results of the feature ablation experiments using LR model are shown in Ta-

Dataset	Method	Precision	Recall	F1
MCTest	BOW	0.41	0.50	0.45
	Unstruct-LR	0.71	0.63	0.67
	LSNM	0.70	0.84	0.74
	Discourse-LR	**0.63**	**0.83**	**0.71**
	SkipTh-BOW	**0.72**	**0.68**	**0.70**
	SkipTh-ALL	**0.70**	**0.84**	**0.76**
Simple Wiki	BOW	0.28	0.20	0.23
	Unstruct-LR	0.50	0.09	0.15
	LSNM	0.38	0.21	0.27
	Discourse-LR	**0.32**	**0.82**	**0.46**
	SkipTh-BOW	**0.71**	**0.26**	**0.38**
	SkipTh-ALL	**0.33**	**0.16**	**0.22**

Table 6.6: Results of previous work and **our models** for the Fulfilled class, on MCTest and SimpleWiki datasets

ble 6.5. The ALL feature set includes all the features described in Section 6.1.1 (without BOW). We obtained high precision and F-measure using the Discourse features. We also experimented with our top feature from the InfoGain analysis, *But-Present*, which surprisingly achieves a high F-measure, compared to using ALL and Discourse feature sets. The last row of Table 6.5 shows the results of using ALL features excluding *But-Present*. This indicates that features motivated by narrative structure are primarily driving improvement. Previous work (Chaturvedi et al., 2016) shows that a model representing narrative structure could beat the BOW baseline, our results suggest that ultimately, the presence of *"but"* is likely a central driver for their improvements as well.

6.1.4 Comparison to Previous Work

We directly compare our methods and features to the most relevant previous work that applied their models on two datasets and reported the results for the Fulfilled class (Chaturvedi et al., 2016) . We present the same metrics in Table 6.6, using our best model **Skip-Thought**

SimpleWiki
DESIRE: wanted INSTANCEID: 5 LABEL: notfulfilled TEXT: * * * The people in Australia * * * # # # wanted to # # # run their own country , and not be told what to do from London . NEXT: The first governments in the colonies were run by Governors chosen by London . Soon the settlers wanted local government and more democracy . The New South Wales Legislative Council , was created in 1825 to advise the Governor of New South Wales , but it was not chosen by voters . William Wentworth established the Australian Patriotic Association (Australia 's first political party) in 1835 to demand democratic government for New South Wales . In 1840 , the Adelaide City Council and the Sydney City Council were started and some people could vote for them (but only men with a certain amount of money).
MCTest
DESIRE: wanted INSTANCEID :1 LABEL: fulfilled TEXT: But after looking at everything , * * * she * * * # # # wanted to # # # get an ice cream sandwich. NEXT: She got the ice cream sandwich. She bit into it and smiled. It tasted so good. She felt so happy. Her brother , Tony , was happy too.

Figure 6.4: Examples of annotated data by Chaturvedi et al. (2016) from SimpleWiki and MCTest datasets, where desire subject and verb are marked by * and # respectively

(SkipTh). We also present results of our LR model with our Discourse features, **Discourse-LR**, trained and tested on their corpora to compare to their features. The first three rows show the results from Chaturvedi et al. (2016) for comparison. They used BOW as baseline, LSNM is their best-performing model, and Unstruct-LR is the unstructured model that uses all of their features with Logistic Regression classifier.

On both corpora, Discourse-LR outperforms Unstruct-LR, showing that the Discourse features are stronger indicators of the desire fulfillment status when used with Logistic Regression classifier. The performance overall is lower on SimpleWiki compared to MCTest dataset. Figure 6.4 shows an example of each corpus, annotated in previous work (Chaturvedi et al., 2016). The stories in MCTest are written for children with the goal of being understandable by

a 7-year old and thus have simpler structure. Understanding such narratives does not require complex features and inferences about the discourse relations in text. The evidence of fulfillment is often explicitly expressed in the context, like *"She got the ice cream sandwich"* and supported by further descriptions (*"She bit into it and smiled. It tasted so good. She felt so happy."*).

On the other hand, SimpleWiki samples are extracted from Wikipedia article with more complexity in sentence structure. For example, in the narrative in Figure 6.4, the desire is to *"run their own country"* and it was not satisfied as we read in the next sentences that the councils and governments were not chosen by voters and by 1840 only a few rich men could vote. Recognizing the unfulfillment of the desire here requires deep semantic understanding to infer that not being able to vote means not running their country. In addition, on SimpleWiki, LR-Discourse outperforms the structured model, LSNM (0.46 vs. 0.27 on F-1), suggesting that our model is a better fit for understanding narratives with more complex structures.

6.2 Further Linguistic Analysis of Thwarted Goals

Interestingly, our experiments and annotation analysis show that identifying the unfulfillment of a goal is a harder task compared to modeling desire fulfillment, suggesting that modeling *Thwarted Goals* could be a challenging and novel problem to explore. An example of a thwarted goal expression is in Figure 6.1: *"I had hoped to ask him to join me for a drink or something after the show (if my courage would allow such a thing) but he left before the end and I didn't see him after that"*. As described in Chapter 3, we created a dataset, *ThwartDB*, includ-

Xcomp	Count	Xcomp	Count
wanted to see	304	wanted to cry	86
wanted to buy	202	wanted to leave	77
wanted to tell	186	wanted to call	77
wanted to say	171	wanted to sleep	69
wanted to ask	122	wanted to work	39
wanted to talk	113	wanted to punch	29

Table 6.7: Examples of frequent clausal complements of *wanted to* in the dataset

Goal Expression Clause	Count
wanted to go	1,217
wanted to go home	35
wanted to take	372
wanted to take picture/pic/photo	127

Table 6.8: Examples of light verb clausal complements with their direct objects

ing $\sim$ 8K expressions of unfulfilled goal expressions. We perform two studies on this dataset. First, we categorize our data to identify common patterns of thwarts, and second, we perform sentiment analysis on the goals and their thwarts to explore whether they can be characterized by the plot unit structure.

6.2.1 Modeling Thwarts

We perform a set of primary studies on ThwartDB dataset using NLP tools. To automatically sort the data based on different types of the goal-expressions, we parsed sentences using Stanford CoreNLP (Manning et al., 2014) and extracted the first open clausal complement (xcomp) of *wanted* verb in each sentence. Table 6.7 shows examples of the frequent clausal complements in the data. For the light verbs like *take* and *get* we also sorted the data by the direct object (dobj) of the clausal complement verb. Table 6.8 shows examples of frequent light verb xcomp and examples of their frequent direct objects. We collapsed direct

Root	Count	Root	Count
not want	224	afraid	67
not know	139	forget	45
could not	142	tired	34
not find	92	busy	31
not sure	41	late	31
not feel	35	miss	20
not able	28	refuse	19

Table 6.9: Examples of roots of thwart-expressions in the dataset

Thwart Expression Root	Count
have	702
not have	222
not have time	50
not have money/cash	16
not have chance/opportunity	11

Table 6.10: Examples of light verbs (have) in thwart-expression clauses

objects semantically when possible and merged them in one category, for example *take picture/pic/photo/photograph*.

To sort the thwart-expressions, we parsed each clause (that follows *but*) and extracted the root of the dependency parse tree with its dependency relations. For example, consider the following sentence:

*I **wanted to** buy a bag, **but** I couldn't find good one.*

The goal-expression is *I wanted to buy a bag* and the thwart-expression is *but I couldn't find good one*. The clausal complement (xcomp) of *wanted* is **buy**, and the root of the thwart-expression clause is **find** which has been negated (**not find**). Table 6.9 shows examples of frequent thwart clause roots. From thwart-expressions in the dataset, about 26% are negated, so we counted negated cases separately, for example *not want* and *not know* in Table 6.9. Similar to goal-expressions, we extracted direct object (dobj) from the dependency parse for the light verbs. One of the most common light verbs among the thwart clause root is

Expression	# Very Positive	# Positive	# Neutral	# Negative	# Very Negative
Goal	6	1,099	6,042	1,902	16
Thwart	7	767	1,575	6,610	89

Table 6.11: Distribution of sentiment labels in goal and thwart expressions

verb *have*. Table 6.10 shows frequency of cases where the root of the thwart-expression is the verb *have* with some of its common direct objects. While counting, we merge the direct objects that are semantically close, like *money/cash* and *chance/opportunity*.

6.2.2 Sentiment Analysis

We explore whether modeling the change of sentiment between the *goal* and its *thwart*, motivated by plot units model. We run the Stanford CoreNLP Sentiment Tool (Socher et al., 2013), which labels every sentence with one of the five labels: Neutral, Positive, Very Positive, Negative, Very Negative. Table 6.11 shows the distribution of sentiment labels for goal and thwart expressions in our dataset. We also examined the sentiment transition between the goal-expression and the thwart-expression for Positive (including Very positive) and Negative (including Very negative) labels, illustrated in Table 6.12. We present some examples of the Positive $\rightarrow$ Negative sentiment transition from Thwarted Goals data in Figure 6.5, where the goal-expression has a Positive (or Very positive) sentiment label and the thwart-expression is Negative (or Very negative).

The Thwarted Goals dataset includes the Stanford CoreNLP annotations as described for each sentence, such as the *xcomp* of the *wanted* verb and all of its dependency relations generated by the parser. The dataset also includes the root of the thwart-expression for each

Goal-Sentiment → Thwart-Sentiment	Count
Positive → Positive (PP)	89
Positive → Negative (PN)	822
Negative → Positive (NP)	186
Negative → Negative (NN)	1,397

Table 6.12: Sentiment transitions subtypes

I really wanted to be strong for everyone else in my family and for garys wife, but I lost it.
I always wanted to be a police-woman but was told I didn't meet the height requirement when I graduated from high school.
I wanted to speak all the witty and smart things that ran through my head, but I couldn't find the words to say them.
I really wanted to tell the women how great they looked and take their photo, but it was so difficult to communicate and you could just tell they were sick of tourists with cameras.
I adored this play when we studied it at school, I always wanted to see it but missed it.
Oh yeah, we wanted to go bowling with Brianna on Thursday after I got off work, but the bowling alley was closed because their power was 'off' (not necessarily 'out').
I wanted to buy a print of that beautiful picture but it turned out to be so expensive - 30 or 40.

Figure 6.5: Examples of transition from Positive to Negative sentiment in Thwarted Goals dataset

sentence with its dependency relations. In addition, each goal-expression and thwart-expression is given a sentiment-label generated by the Stanford Sentiment Analyzer. The dataset and studies presented in this section aim to provide a foundation and testbed for modeling the thwarted goals in narratives and motivate researchers for future work in this area.

6.3 Summary

Previous work has attempted to develop computational models of affects showing that it is a difficult problem given the performance of the current NLP tools, even on small sets of well-structured and short stories, with manual annotations and simplified assumptions. The

final part of our work presented in this Chapter is focused on recognizing the expression of the protagonist's goals and desires and tracking their corresponding outcomes, as a sub-problem of modeling affects. We develop and test our methods on open-domain personal blog stories and show that we can identify the fulfillment status of the desires with an overall F-1 score of 0.7.

We introduce a new corpus, DesireDB, consisting of desire expressions with their context and annotations of the *desire*, its *fulfillment status*, and the *evidence*. Our studies show that both prior and post context are useful for modeling desire fulfillment. Moreover, we show that exploiting narrative structure is helpful, both directly in terms of the utility of discourse relation features and indirectly via the superior performance of a Skip-Thought LSTM model.

We further studied the goal expressions by creating an additional corpus including $\sim$ 8K expressions of unfulfilled goals. Our experiments indicate that the sentiment transitions in unfulfilled goals do not always follow the failure plot structure and only about 24% of our data contains thwarts with negative polarity. This result suggests that other conceptual structures are needed to provide a comprehensive model of the goals and their outcomes and also that off-the-shelf sentiment analyzers perform poorly on this genre of narratives.

Overall, our studies suggest that several challenging problems remain unsolved to reliably identify and model affective states in narratives. We explored one of the implicit affect expressions in stories, the goals and desires of the protagonists. There are several other ways that affects can be expressed in the narratives, that need to be modeled. For example implicitly in the events that occur (*"I had to go to the emergency room last night"*) and explicitly expressing emotions (*"I was happy to see them"*).

As for goal expressions, we have focused on the verbal patterns of goals, however,

humans can express their desires in many different ways, e.g. using adjectives in a stative sentence: *"I was thirsty"*. We propose studying other linguistic frames of desire expressions as a future work for modeling the narrator's goals. We have used high-precision search patterns but our annotations show that such patterns still admitted some hypothetical desires (e.g., "If I had wanted to buy a book"). It would appear that distinguishing hypothetical vs. real desires itself is another interesting problem to tackle. DesireDB contains annotator-labeled spans for evidence for the annotator's conclusions. While we have not used this labeling, we believe that using the evidence is worth exploring in future.

Chapter 7

Conclusions and Future Work

As stories play an important role in human understanding and perception of the world, the computational analysis of narrative structure is a key area in natural language processing research. This thesis presents our work on computational models of events and desires in everyday lives of humans, using a corpus of personal narratives written by ordinary people on the blogs. Our unsupervised approach for modeling events generates contingent pairs of fine-grained events from datasets of blog stories on different topics. We identify desires of the narrators expressed in their stories through explicit verbal patterns and predict their outcome, whether they accomplished their goal or not.

We describe our contributions in Section 7.1. The limitations of our framework is described in Section 7.2 along with suggestions for short-term future work to address them. We describe four different directions for future work, such as generating narrative schema and narrative comprehension models in Section 7.3. We conclude by summarizing our work in Section 7.4.

7.1 Contributions

We summarize the main contributions of this thesis in three parts.

7.1.1 Linear Model of Narrative Structure

Our first contribution is developing a supervised method for automatically identifying a high-level linear structure of the narratives based on Labov and Waletzky's theory of oral narratives. We created a dataset with gold-standard labels based on three categories of narrative clauses: Orientation, Action, Evaluation, with inter-annotator agreement kappa of 0.63. Our best classifier achieved an F-1 measure of 0.69, which was increased to 0.77 in the experiments on the data points where all three annotators agreed on the labels.

We studied some properties of the data in our experiments and measured performance for each story rather than over clauses. Our results indicated that some stories are more difficult to model than others and that ambiguous clauses are likely to be correlated with particular writing styles. This suggests that different genres of narrative data could be challenging in different ways and requiring different methods.

Moreover, we developed a classifier to distinguish actions from non-action clauses, achieving an F-1 score of 95% overall and precision of 90% for the action category. We had hypothesized that reliably identifying action clauses in the narratives and using them for modeling events can enhance the methods for learning contingency relations between events. However, our results indicated no improvement when using actions only (using our classifier results) compared to using the entire narratives. Therefore, in the next two parts of our contributions we

proposed other methods for modeling events and affects as the main elements of the narrative structure.

7.1.2 Modeling Events and their Contingency Relations

We presented an unsupervised method for learning contingency relation between events from personal narratives and directly compared our methods to several other approaches as baselines. We used a weakly-supervised algorithm to identify the indicative events from domain-specific personal narratives and applied a bootstrapping approach using these event patterns to create topic-sorted datasets of blog stories. Moreover, we proposed new evaluation methods for testing the contingent event pairs, using human judgments and automatic two-choice questions. The evaluations show that 82% of the relations between events that we learn from topic-sorted stories are judged as contingent. Using our models, we generated collections of contingent event pairs, both topic-specific and domain-independent.

We used two different datasets to directly compare what is learned from topic-sorted stories as opposed to a general-domain story corpus. Our results show that using a topic-specific corpus where the stories share the same theme could reveal more fine-grained event relations. The results show significant improvement over the Event-Unigram, Event-Bigram, and Event-SCP (Rel-grams method) baselines on topic-specific stories: 25% improvement of accuracy on Camping Trip and 41% on Storm topic compared to Bigram model (top-performing baseline). Our evaluations indicate that a method that works well on the news genre does not generate coherent results on personal stories (comparison of Event-SCP baseline with Causal Potential). Moreover, we show that most of the fine-grained contingency relations we learn from narra-

tive events are not found in existing narrative and event schema collections induced from the newswire datasets (Rel-grams).

7.1.3 Modeling Desires and their Outcomes

In the final part of our work we developed supervised methods for modeling the protagonist's goals in general-domain first-person narratives. We trained and tested our methods on open-domain personal blog stories and show that we can identify the fulfillment status of the desires with an overall F-1 score of 0.7. Our studies show that both prior and post context are useful for modeling desire fulfillment. In addition, we show that exploiting narrative structure is helpful, both directly in terms of the utility of discourse relation features and indirectly via the superior performance of a Skip-Thought LSTM model.

We created a corpus of desire expressions, *DesireDB*, with prior and post narrative context including gold-standard annotations for Fulfillment Status, Evidence of Status, and Annotator-Agreement Scores. Construction of this corpus involves a systematic method for identifying and extracting goal expressions in personal narratives. Moreover, we created a dataset of thwarted goals, *ThwartDB*, including unaccomplished desire expressions and presented our primary studies on common patterns of thwarts and their sentiment transitions. Our experiments indicate that the sentiment transitions in unfulfilled goals do not always follow the failure plot structure and only about 24% of our data contains thwarts with negative polarity. This result suggests that other conceptual structures are needed to provide a comprehensive model of the goals and their outcomes and also that off-the-shelf sentiment analyzers perform poorly on this genre of narratives.

7.2 Limitations

We summarize the major limitations of our frameworks and methods in three categories in this Section. Investigating each item described below can serve as a direction for future work in the short term.

1. The linear model of narrative structure based on L&W's theory

    - The sentences in the narratives can have multiple functions and fit into more than one category of L&W. This limitation suggests that their framework may be too high-level for separating the key elements of narrative. Other theories that propose more fine-grained structures for stories could be explored as alternatives.
    - Isolating one part of narrative structure, such as actions, does not enhance the methods developed for modeling narrative events. Different elements of narrative structure may need to be modeled in relation to each other. For example, some of the affect states of the protagonists are implicit in the expression of events and thus it is not possible to model affects without considering the events timeline. Exploring other models that focus on the relation between different parts of narratives, such as Plot Units, could be the next step to overcome this limitation.

2. Modeling events and their contingency relation

    - We define an event structure as a verb with its arguments, however, events can be expressed in different ways such as nouns, e.g. *explosion*. A follow-up on our work could study other forms of event expressions in narratives.

- Our work did not construct a complete timeline of the narrative events. However, we believe that the contingent event pairs could potentially be used to build longer chains of causally related events.
- A detailed error analysis is needed to show why Causal Potential works better than previous methods such as PMI, Bigram and SCP.
- Our results indicated that the methods that work well on the news genre do not generate similar results on personal blog stories. An analysis of different corpora is required to identify the type differences between narrative genres and the finding out the reasons why various methods work differently on each dataset.
- We focus on unsupervised models and another interesting future work is to explore using of supervised methods for learning causality between events. We propose that the top contingent event pairs generated using our methods (with high confidence scores) can be used as gold-standard for supervised algorithms.

3. Modeling goals and their outcome

- We studied the protagonist desires which are one way of implicitly expressing affects in the narratives. However, there are various other ways that affects can be expressed in the narratives. As a follow-up on our approach we propose exploring other types of affect expressions in the narratives such as explicit statements of emotions.
- Our work only used the verbal patterns of desires, however, they can be expressed using other linguistic frames such as adjectives. Generating other patterns of desire

expressions can result in creating larger and more diverse corpora for modeling narrative goals.

- We identify and model the desires that are explicitly stated in the narrative. Extracting the implicit goal expressions from narratives and modeling them is another problem to be tackled in future.

7.3 Future Work

We discuss four major topics of interesting and challenging problems for further research on modeling narrative events and affects.

1. **Topic Modeling of Personal Blogs.**

   We propose to explore existing topic-modeling algorithms, such as LDA, to create a broader set of topic-sorted corpora. The topic-specific event patterns generated in our work can be used as a seed set for semi-supervised variations of topic models. We posit that such corpora will be a valuable resource for further analysis of narrative structure and generating narrative schema.

2. **Generating Narrative Scripts.**

   The main idea behind generating scripts was to build chains of fine-grained events that form a coherent timeline. We posit that the contingent pairs of fine-grained events that we generated in our research can serve as building blocks for creating coherent narrative schema. However, additional methods are as we discussed in Chapter 5. We propose creating a graph of events, weighted by causal potential scores could have the capacity of

revealing sequences of causally related events.

3. **Identifying Implicit Affects in Event Expressions.**

   The events that occur in a story can have an affective impact on their agents and patients. Previous work has explored creating collections of events with positive and negative effects on the characters (Ding and Riloff, 2016). We posit that this work could be expanded to identify the implicit affects expressed in the narrative events by further analyzing the context surrounding the event expression and the sentiment transition in the context. We propose that further features and classification models can be applied to improve the results of the previous work on this topic.

4. **Narrative Comprehension.**

   We posit that DesireDB can be used for several machine comprehension tasks, in the form of questions with multiple answers, such as:

   - Given the prior context, choose the correct desire expression. The prior context of each data instance in DesireDB will be used as the question, the desire expression is the correct answer, and the other (incorrect) answers are randomly selected from other data instances' desire expressions.
   - Given a desire expression and its fulfillment status, choose the correct evidence. The desire expression of each data instance and its status forms the question and the correct answer is the span of text marked as evidence. The other (wrong) choices will be selected randomly from other data instances or from other parts of context that was not marked as evidence.

We propose that the problems described above provide interesting challenges for developing and testing machine learning methods.

7.4 Summary

Several challenges exist in developing computational models of narrative text as we have discussed throughout this thesis. Our work focuses on tackling a subset of these problems. One of the main differences of our work compared to previous research on modeling narratives is using an open-domain corpus of blog stories. We developed and evaluated models of events and desires expressed in personal narratives and showed that we achieved improvements compared to the previous work.

We learned that a linear high-level model of narratives such as L&W, may not simplify the process of modeling narrative structure as we had initially hypothesized. Our analysis and observations indicate that current NLP technologies, such as sentiment analyzers and parsers, generate several errors when applied to the blogs corpus, probably because they are mainly trained and tested on other genres such as news articles.

Our major contributions are (1) the unsupervised method developed for learning fine-grained contingency relation between events, (2) supervised methods for identifying and modeling desires expressed in narratives, and (3) several corpora and data collections such as contingent event pairs, DesireDB, ThwartDB, and L&W annotations. We conclude our work by proposing several directions of future work, both short term (such as detailed error analysis of our results) and long term (such as the narrative schema and narrative comprehension models).

Bibliography

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *The semantic web*, pages 722–735.

Baker, C., Fillmore, C., and Lowe, J. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Balasubramanian, N., Soderland, S., Mausam, and Etzioni, O. (2013). Generating coherent event schemas at scale. In *EMNLP*, pages 1721–1731.

Bamberg, M. (2006). Stories: Big or small: Why do we care? *Narrative inquiry*, 16(1):139–147.

Barthes, R. (1978). *Image-music-text*. Macmillan.

Bartlett, F. F. C. and Bartlett, F. C. (1995). *Remembering: A study in experimental and social psychology*, volume 14. Cambridge University Press.

Beamer, B. and Girju, R. (2009). Using a bigram event model to predict causal potential. In *Computational Linguistics and Intelligent Text Processing*, pages 430–441. Springer.

Bohanek, J. G., Marin, K. A., and Fivush, R. (2008). Family narratives, self, and gender in early adolescence. *The Journal of Early Adolescence*, 28(1):153–176.

Bruner, J. (1991). The narrative construction of reality. *Critical inquiry*, pages 1–21.

Burton, K., Java, A., and Soboroff, I. (2009). The ICWSM 2009 Spinn3r dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*.

Burton, K., Kasch, N., and Soboroff, I. (2011). The icwsm 2011 spinn3r dataset. In *Proceedings of the Annual Conference on Weblogs and Social Media (ICWSM)*.

Chambers, N. and Jurafsky, D. (2008). Unsupervised learning of narrative event chains. *Proceedings of ACL-08: HLT*, pages 789–797.

Chambers, N. and Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. In *Proceedings of the 47th Annual Meeting of the ACL*, pages 602–610.

Chaturvedi, S., Goldwasser, D., and III, H. D. (2016). Ask, and shall you receive? understanding desire fulfillment in natural language text. In *Proceedings of the National Conference on Artificial Intelligence*.

Chollet, F. (2015). Keras.

Coster, W. and Kauchak, D. (2011). Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:*

Human Language Technologies: short papers-Volume 2, pages 665–669. Association for Computational Linguistics.

Ding, H. and Riloff, E. (2016). Acquiring knowledge of affective events from blogs using label propagation. In *AAAI*, pages 2935–2942.

Do, Q. X., Chan, Y. S., and Roth, D. (2011). Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.

Elson, D. (2012a). Dramabank: Annotating agency in narrative discourse. In *LREC*, pages 2813–2819.

Elson, D. and McKeown, K. (2009). A tool for deep semantic encoding of narrative texts. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 9–12.

Elson, D. and McKeown, K. (2010). Building a bank of semantically encoded narratives. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Malta*. Citeseer.

Elson, D. K. (2012b). Detecting story analogies from annotations of time, action and agency. In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative, Istanbul, Turkey*.

Elson, D. K. (2012c). *Modeling Narrative Discourse*. PhD thesis, Columbia University, New York City.

Fauceglia, N. R., Lin, Y.-C., Ma, X., and Hovy, E. (2015). Word sense disambiguation via propstore and ontonotes for event mention detection. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 11–15.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Feng, S., Kang, J. S., Kuznetsova, P., and Choi, Y. (2013). Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Association for Computational Linguistics (ACL)*.

Fivush, R., Bohanek, J. G., and Duke, M. (2005). The intergenerational self: Subjective perspective and family history. *Individual and collective self-continuity. Mahwah, NJ: Erlbaum*.

Fivush, R. and Nelson, K. (2004). Culture and language in the emergence of autobiographical memory. *Psychological Science*, 15(9):573–577.

Forster, E. M. (2010). *Aspects of the Novel*. RosettaBooks.

Gerrig, R. (1993a). *Experiencing narrative worlds: On the psychological activities of reading*. Yale Univ Pr.

Gerrig, R. (1993b). *Experiencing narrative worlds: On the psychological activities of reading*. Yale Univ Pr.

Girju, R. (2003). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pages 76–83. Association for Computational Linguistics.

Gordon, A., Bejan, C., and Sagae, K. (2011). Commonsense causal reasoning using millions of personal stories. In *Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*.

Gordon, A. and Swanson, R. (2009). Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA*.

Goyal, A. and Riloff, E. (2013). A computational model for plot units. *Computational Intelligence*, 29(3):466–488.

Goyal, A., Riloff, E., and Daumé III, H. (2010). Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 77–86.

Graesser, A. C., Singer, M., and Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological review*, 101(3):371.

Habermas, T. and Bluck, S. (2000). Getting a life: the emergence of the life story in adolescence. *Psychological bulletin*, 126(5):748.

Hall, M., Eibe, F., Holms, G., Pfahringer, B., Reutemann, P., and Witten, I. (2005). The weka data mining software: An update. *SIGKDD Explorations*, 11(1).

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hu, Z., Rahimtoroghi, E., Munishkina, L., Swanson, R., and Walker, M. A. (2013). Unsuper-

vised induction of contingent event pairs from film scenes. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 370–379.

Joskowicz, L., Ksiezyck, T., and Grishman, R. (1989). Deep domain models for discourse analysis. In *AI Systems in Government Conference, 1989., Proceedings of the Annual*, pages 195–200. IEEE.

Kaplan, R. M. and Berry-Rogghe, G. (1991). Knowledge-based acquisition of causal relationships in text. *Knowledge Acquisition*, 3(3):317–337.

Khoo, C. S., Chan, S., and Niu, Y. (2000). Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 336–343. Association for Computational Linguistics.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. *Sociological methodology*, 2:139–150.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Sage.

Labov, W. (1972). The transformation of experience in narrative syntax. In *Language in the Inner City*, pages 354–396. University of Pennsylvania Press, Philadelphia.

Labov, W. (1997). Some further steps in narrative analysis. *Journal of narrative and life history*, 7:395–415.

Labov, W. and Waletzky, J. (1967). Narrative Analysis: Oral Versions of Personal Experience. *Journal of Narrative and Life History*, 7(1-4):3–38.

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

Louis, A. and Nenkova, A. (2011). Automatic identification of general and specific sentences by leveraging discourse annotations. In *IJCNLP*, pages 605–613.

Magliano, J. P. and Radvansky, G. A. (2001). Goal coordination in narrative comprehension. *Psychonomic Bulletin & Review*, 8(2):372–376.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Manshadi, M., Swanson, R., and Gordon, A. S. (2008). Learning a probabilistic model of event sequences from internet weblog stories. In *Proceedings of the 21st FLAIRS Conference*.

Maslow, A. H. (1943). A theory of human motivation. *Psychological review*, 50(4):370.

Max-Neef, M., Elizalde, A., and Hopenhayn, M. (1992). Development and human needs. *Real-Life Economics: understanding wealth creation, edited by: Ekins, P. and Max-Neef, M., Routledge, London*, pages 197–213.

McAdams, D. P., Josselson, R. E., and Lieblich, A. E. (2006). *Identity and story: Creating self in narrative.* American Psychological Association.

McLean, K. C. and Thorne, A. (2003). Late adolescents' self-defining memories about relationships. *Developmental Psychology*, 39(4):635.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miller, G. and Fellbaum, C. (1998). Wordnet: An electronic lexical database.

Miller, P. J., Fung, H., and Koven, M. (2007). Narrative reverberations: How participation in narrative practices co-creates persons and cultures.

Nguyen, K.-H., Tannier, X., Ferret, O., and Besançon, R. (2015). Generative event schema induction with entity disambiguation. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics (ACL-15)*.

Ouyang, J. and McKeown, K. (2014). Towards automatic detection of narrative structure. In *LREC*, pages 4624–4631. Citeseer.

Ouyang, J. and McKeown, K. (2015). Modeling reportable events as turning points in narrative. In *EMNLP*, pages 2149–2158.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English gigaword. *Linguistic Data Consortium*.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *LIWC: Linguistic Inquiry and Word Count*.

Pichotta, K. and Mooney, R. J. (2014). Statistical script learning with multi-argument events. *EACL 2014*, page 220.

Poynor, D. V. and Morris, R. K. (2003). Inferred goals in narratives: Evidence from self-paced reading, recall, and eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(1):3.

Pratt, M. W. and Fiese, B. H. (2004). *Family stories and the life course: Across time and generations*. Routledge.

Rahimtoroghi, E., Corcoran, T., Swanson, R., Walker, M. A., Sagae, K., and Gordon, A. S. (2014). Minimal narrative annotation schemes and their applications. In *7th Workshop on Intelligent Narrative Technologies*, Milwaukee, WI.

Rahimtoroghi, E., Hernandez, E., and Walker, M. A. (2016). Learning fine-grained knowledge about contingent relations between everyday events. In *Proceedings of SIGDIAL 2016*, pages 350–359.

Rahimtoroghi, E., Swanson, R., and Walker, M. A. (2013). Evaluation, orientation, and action in interactive storytelling. In *Proceedings of Workshop Intelligent Narrative Technologies 6*.

Rapp, D. and Gerrig, R. (2006). Predilections for narrative outcomes: The impact of story contexts and reader preferences. *Journal of Memory and Language*, 54(1):54–67.

Reed, L., Wu, J., Oraby, S., Anand, P., and Walker, M. (2017). Learning lexico-functional patterns for first-person affect. In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics (ACL-17)*. ACL.

Riaz, M. and Girju, R. (2010). Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, pages 361–368. IEEE.

Riaz, M. and Girju, R. (2013). Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *SIGDIAL Conference*, pages 21–30.

Richards, E. and Singer, M. (2001). Representation of complex goal structures in narrative comprehension. *Discourse Processes*, 31(2):111–135.

Richardson, M., Burges, C. J., and Renshaw, E. (2013). Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, volume 3, page 4.

Riloff, E. (1996). Automatically generating extraction patterns from untagged text. In *Proceedings of the national conference on artificial intelligence*, pages 1044–1049.

Riloff, E. and Phillips, W. (2004). An introduction to the sundance and autoslog systems. Technical report, Technical Report UUCS-04-015, School of Computing, University of Utah.

Roemmele, M., Bejan, C. A., and Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Schank, R., Abelson, R., and Schank, R. C. (1977). *Scripts Plans Goals*. Lea.

Slater, M. D. and Rouner, D. (2002). Entertainmentducation and elaboration likelihood: Understanding the processing of narrative persuasion. *Communication Theory*, 12(2):173–191.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Swanson, R., Rahimtoroghi, E., Corcoran, T., and Walker, M. A. (2014). Identifying narrative clause types in personal stories. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Tetreault, J., Chodorow, M., and Madnani, N. (2013). Bucking the trend: improved evaluation and annotation practices for esl error detection systems. *Language Resources and Evaluation*, pages 1–27.

Thorne, A. (2004). Putting the person into social identity. *Human Development*, 47(6):361–365.

Thorne, A., Korobov, N., and Morgan, E. M. (2007). Channeling identity: A study of storytelling in conversations between introverted and extraverted friends. *Journal of research in personality*, 41(5):1008–1031.

Thorne, A. and McLean, K. C. (2003). Telling traumatic events in adolescence: A study of master narrative positioning. *Connecting culture and memory: The development of an autobiographical self*, pages 169–185.

Thorne, A., McLean, K. C., and Lawrence, A. M. (2004). When remembering is not enough: Reflecting on self-defining memories in late adolescence. *Journal of personality*, 72(3):513–542.

Thorne, A. and Shapiro, L. A. (2011). Testing, testing: Everyday storytelling and the construction of adolescent identity. *Adolescent Vulnerabilities and Opportunities: Developmental and Constructivist Perspectives*, 38:117.

Trabasso, T. and van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language*, 24:612–630.

Wiebe, J. M. (1990). Identifying subjective characters in narrative. In *Proceedings of the 13th conference on Computational linguistics-Volume 2*, pages 401–406. Association for Computational Linguistics.

Wilensky, R. (1982). Points: A theory of the structure of stories in memory. In Lehnert, W. G. and Ringle, M. H., editors, *Strategies for Natural Language Processing*.