Socially-Aware Machine Learning:

Towards Leveraging the Relationship Between Narrative Comprehension and Mentalizing

by Prashanth Vijayaraghavan

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June, 2021

Author .....

Program in Media Arts and Sciences

May 21, 2021

Certified by .....

Deb Roy

Professor of Media Arts and Sciences

Thesis Supervisor

Accepted by .....

Tod Machover

Academic Head

Program in Media Arts and Sciences

Socially-Aware Machine Learning:
Towards Leveraging the Relationship Between Narrative Comprehension and Mentalizing

by Prashanth Vijayaraghavan

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on May 21, 2021, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

ABSTRACT

Narratives are the fundamental means by which people organize, understand, and explain the social world. Research suggests that exposure to narratives improves mentalizing, referring to the capacity to forecast and reason about others' mental states. Simultaneously, enhanced mentalizing abilities are closely linked to exhibiting improved narrative processing skills. The purpose of this dissertation is to develop modular computational methods that leverage the relationship between mentalizing and narrative comprehension for understanding specific aspects of social-cognitive processes and seek to advance the research towards imparting social awareness to machines. Our work consists of three main functional modules. First, we present a representation learning approach that computes a social situational embedding of sentence-level social events. Next, we apply the learned social event representation to embed, infer and explain the characters' mental states from the narratives. Finally, we analyze some of the basic elements of narrative structure present in short personal narratives as a means of exemplifying the story understanding capability. Particularly, we investigate the role of characters' cognitive tension captured using our inferred mental representation for automatically detecting the central conflict of a story i.e. the climax and their resolution.

Unlike most previous work that either uses conventional trait-based models or exploits low-level annotations of short fixed-length stories, we tackle a subset of the data and modeling challenges directed at inferring human motives and emotional reactions. First, we construct a relatively open-ended corpus of personal narratives and commonsense knowledge from social media containing more variations in terms of topical content. Using this weakly annotated corpus, we train deep learning models that compute rich representations of social events capturing aspects of syntactic, semantic, and pragmatic properties and integrate them to generate textual explanations of motives and emotions of characters in the narrative. Empirically, our proposed approaches outperform several baselines in mental state tracking tasks and harness transferability to low-resource regimes and other downstream tasks.

As a final contribution in this dissertation, we demonstrate improved narrative processing skills by computationally predicting key elements of narrative structure in personal narratives. Notably, our studies show that integrating the protagonist's mental state embeddings with linguistic information leads to the enhanced prediction of climax and resolution in narratives. Our data and modeling contributions emphasize the value of exploiting the mutual influence of mentalizing and narrative comprehension, thereby promoting future efforts towards building human-centered AI systems.

Socially-Aware Machine Learning:

Towards Leveraging the Relationship Between Narrative Comprehension and Mentalizing

by Prashanth Vijayaraghavan

This doctoral thesis has been reviewed and approved by the following committee members:

Deb Roy .....

Thesis Committee Chair

Professor

MIT Media Lab

David Bamman .....

Thesis Reader

Assistant Professor

University of California, Berkeley

Douwe Kiela .....

Thesis Reader

Research Scientist

Facebook

Chandra Bhagavatula .....

Thesis Reader

Research Scientist

Allen AI

"Cultivation of mind should be the ultimate aim of human existence. "

B.R.Ambedkar

Acknowledgements

My grad school journey has been a great learning experience and I feel fortunate to have many wonderful people around to help me see through it. First and foremost, I would like to express my gratitude to my advisor Prof. Deb Roy, whose expertise, understanding, and patience, added greatly to my experience at the lab. Deb has always provided the freedom and support to work on projects that I found interest in. I sincerely thank him for all the opportunities, guidance and support during the time of research and writing of this dissertation. I would like to thank my committee members, Chandra Bhagavatula, David Bamman and Douwe Kiela for their guidance, insightful comments, and feedback throughout this process. I feel extremely fortunate to have been able to interact and work with them.

Special thanks to Heather Pierce and Keyla Gomez for being really concerned and supportive through out the Ph.D. program. My journey would have been incomplete without my amazing friends and colleagues. I want to thank all of them for the stimulating discussions and all the fun we have had together in the lab: Soroush Vosoughi, Brandon Roy, Misha Sra, Mina Soltangheis, Sneha Priscilla Makini, Eric Chu, Nabeel Gillani, Martin Saveski, Juliana Nazare, Anneli Hershman, Bridgit Mendler, Belen Saldias, Ivan Sysoev, Lauren Fratamico, Ann Yuan, Shayne O'Brien, William Brannon, Luke, Maggie Hughes, Nazmus Saquib, Mark Exposito, Neo Mohsenvand, Eric Pennington, Perng-hwa Kung, Alex Siegenfeld, Sophie Chou, Hang Jiang, Suyash Fulay, Sarah Ballinger, Wonjoune Kang, David McClure, Doug Beeferman, Russell Stevens, William Powers, Andrew Heyward and Preeta Bansal.

Last, but certainly not least, I thank whole of my family, my childhood friends, friends from school, UG, past jobs and LSM for their continued support and encouragement through my years at MIT. My love for them is unfathomable.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Ubiquity and Importance of Narratives

> *"The narratives of the world are numberless. Narrative is first and foremost a prodigious variety of genres, themselves distributed amongst different substances as though any material were fit to receive mans stories. Able to be carried by articulated language, spoken or written, fixed or moving images, gestures, and the ordered mixture of all these substances; narrative is present in myth, legend, fable, tale, novella, epic, history, tragedy, drama, comedy, mime, painting (think of Carpaccios Saint Ursula), stained glass windows, cinema, comics, news item, conversation."*
>
> BARTHES [BAR66; BD75]

Narratives are one of the most common yet powerful means of communication used to enhance engagement with people's issues and understanding of the social world. Psychologist Robyn Dawes [Daw99] even claimed that humans are "the primates whose cognitive capacity shuts down in the absence of a story". Regardless of ethnicity, language, and enculturation, the ubiquity of narratives has been emphasized in several interdisciplinary academic studies, including literary studies, anthropology, sociolinguistics, psychology, artificial intelligence, to list a few [Cha80]. This is reflected in

Barthes' analysis of narrative [BD75]. Because of the pervasiveness of narrative in our lives, it is viewed as an essential aspect of making sense of one's own experiences, underpinning everyday thinking and expression [SEG01]. The themes and characters in these stories reflect real-world conflicts, solutions, humor, cultural values, people's psychological states, and their personality. Bruner [Bru09] argued that social cognition is mediated by a specialized mode, referred to as the "narrative mode", which functions by granting causal efficacy onto psychological states in determining the behavior of the self and others, an idea that explains folk psychology[Hut07; Hut12]. Moreover, stories tend to form the basis of our memories, thoughts, and knowledge as information stored as a narrative is quickly interpreted and better remembered than those that are organized into non-narrative frameworks [GOK; WJ14]. In this work, we use the terms narratives and stories interchangeably.

As stories present different facets of the social world, humans constantly hypothesize and represent mental states of the various interactants (or participants) in social situations according to their social actions [SK03; BST09]. This cognitive capability to infer and represent one's own and other people's mental states (e.g., motivations, emotions, thoughts, beliefs, desires, and attitudes) is referred to as mentalizing or theory-of-mind (ToM). Beyond interpreting the social world, this complex of abilities enables us to understand prosocial behavior such as to empathize, build peer relationships, form judgments, provide care, to name a few [Eis14; WM08; Imu+16]. This mentalizing network is considered to be common to social cognition and narrative comprehension involving ToM as a common component in which people simulate the mental states of other people (or characters). Mentalizing promotes the construction of a mental model of the story and is deemed essential for enhanced story comprehension capability. On the flip side, narratives play a potent role in inferring mental states and personality traits with the help

of social knowledge related to the real-world, evoking the social-cognitive mechanisms. Repeated evocation through regular engagement with stories contributes to the improvement of social cognition [MO08; MJ09]. All this has led several theorists and researchers to suggest how stories exert a powerful influence on social cognition and vice versa [Mar18; Oat99; Hog03].

1.2 Research Goal

Relating Mentalizing and Narrative Comprehension

A growing body of work has been developed in neuroscience concerning the interdependence between mentalizing and narratives, [GW02; GPHL08; CSG98; CWC11]. Several neuroimaging studies have emphasized how the overlapping brain regions implicated for both story comprehension, and mentalizing reflect their interdependence – (a) narratives act as an instrument in the evocation of mentalizing processes, i.e., reading more stories in one's lifetime and analyzing characters' behavior in stories contribute to greater activation of mentalizing network [Fer+08; Mar11; Tam+16b; MG17], and (b) enhanced mentalizing ability is closely linked to exhibiting improved narrative processing skills owing to the necessity of inferring characters' mental states and understanding their complex social relations present in narratives [Mar+06; AP; Fer13; McK92; LT18; Kim14]. This interdependence is often linked to the acquisition of an articulated mental language, comprising elaborate social and emotional vocabulary. Although the mutual influence of mentalizing and narrative comprehension has received the most empirical attention in the field of neuroscience, it is still a topic open to much more investigation in AI research and, indeed, essential to impart social intelligence to machines.

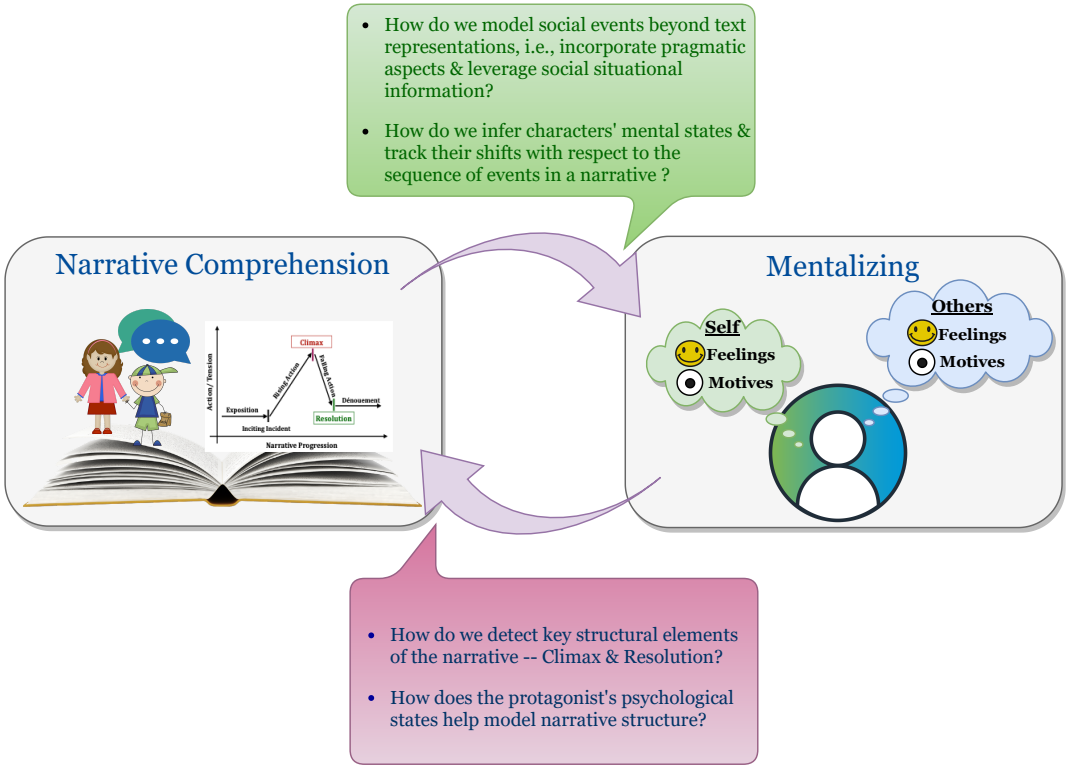


FIGURE 1.1: Overview of our goal: From modeling each social event through a pragmatic lens to inferring trajectories of characters' mental states over sequences of events in narratives to fleshing out details of high-level structural components of the narrative, we leverage the influence between narrative comprehension and mentalizing (or ToM) and demonstrate how they contribute towards building improved AI systems.

The primary goal of this dissertation is to develop modular computational models that will leverage the interdependence between mentalizing and narrative comprehension to model specific aspects of socio-cognitive processes and further the research towards building AI systems with social awareness. The interplay between them is reflected in the process of storytelling where:

- Narrators present a story with a specific sequence of events necessitating the social and situational interpretation of events.
- Perceivers construct a unified, coherent representation of the story by drawing inferences about the mental states of people (or intentionality

of participants) considering their social interactions and the chronological relationship between event sequences.

- Perceivers make sense of the narrative by deriving the theme and recognizing episode boundaries that represent the high-level story structure.

Figure 1.1 provides an illustration of our overarching goal. We capture this mutual influence between narrative comprehension and mentalizing by focusing on the following functional modules:

- Developing a computational model that produces social situational representation of events that can dispel the ambiguities in the interpretation of the text using social commonsense knowledge.

- Constructing mental models of characters in stories by integrating the learned representation of events computed using social commonsense knowledge and dynamically tracking the shifts in their mental states related to the events in the narrative.

- Demonstrating the ability of these mental state models to promote efficient identification of key elements of the narrative structure like climax and resolution. Integration of psychological states of the protagonist beyond linguistic information could potentially enhance the story understanding capacity.

Combining all these three functional modules that introduce new resources and modeling approaches, we believe our goal would facilitate technological advancements and research in a variety of disciplines that are either interested in developing human-centered social cognition or likely to benefit from such capabilities.

1.3 Characteristics of Narrative Data

Narrative data comprise a diverse range of texts, including novels and short stories, prose and poetry, movie screenplays and synopsis, personal experiences and life stories, oral memoirs, autobiographies and histories. Such diverse texts are considered a narrative if they contain an organized sequence of causally linked events with a meaningful consequence [Rie93]. This description of the narrative is well-aligned with Bruner's definition of narrative:

> *Perhaps its principal property is its inherent sequentiality: a narrative is composed of a unique sequence of events, mental states, happenings involving human beings as characters or actors. These are its constituents. But these constituents do not, as it were, have a life or meaning of their own. Their meaning is given by their place in the overall configuration of the sequence as a whole...*
>
> BRUNER [BRU90]

Personal Narratives

Given the diverse forms of narrative text, we focus on social media stories that are told by common people on a daily basis about their personal experiences. Personal narrative is a form of autobiographical storytelling that gives shape to life experiences. These stories could potentially be unstructured and discursive, yet are essential social resources that (a) build the identities of the tellers and the audiences and (b) learn about the temporal and causal relationships between event sequences. Moreover, personal stories are known to be brief, diverse, and major sources of commonsense causal information, centered around social and mental topics more than natural and physical causality [RBG11; MSG08; GS09].

Several previous studies [Bro+19; Abb20; Bro20] have recognized "focalization" or "protagonism" as a crucial part of storytelling (or narration). It is referred to as the fundamental narrative function of putting forth the perspective of a single person, i.e., the protagonist, in a narrative. While news stories, histories, movies have masses of individuals, personal narratives usually contain few characters told from the protagonist's point of view. By distinguishing the protagonist's perspectives from all other participants, protagonism operates by developing insights into other people as relatable self-proxies even for stories about others [Sto16; DFG11; Lab01]. This means the narrators describe the social world and draw inferences about the characters or participants from their perspective, sometimes even egocentrically related. In the vast majority of stories, they not only express their internal mental states but also impute others' inner states through their stories. Further, this protagonistic approach to the story reflects Bruner's narrative mode of inference based on psychological causation. Hence, while a chronological presentation would explain "how" things happened, we need the protagonist's narrative mode to explain "why" they happened, in other words, what motivated the events (i.e., intentions) from a given personal perspective and how their impact was felt (i.e., emotional reactions) [Bru91b; Bru91a; Bro20]. Together, these properties of personal narratives make them suitable for our research goals. Therefore, we rely on personal narratives obtained from social media as a vital part of our knowledge extraction, modeling, and evaluation phases across various aspects of this dissertation.

Personal narratives, like other stories, have specific plots, sequences of events, and actions ordered to highlight specific themes [Pol88]. They contain a representation of the events comprising entities or characters with specific social roles, group, and cultural context performing planned actions intended to achieve the desired goal state situated within a particular space and time. In addition to interpreting the explicit and implied psychological

states of narrators and other participants in the narrative, there have been theories that identify key structural elements in personal verbal narratives as a result of sociolinguistic studies. A seminal work by Labov and Waletzky [LW97; Lab06] analyzed oral narratives of true personal experience from the narrators' lives, obtained via sociolinguistic interviews. Through this narrative analysis, the significance of basic structural components in narratives [Tho04; TM03; TKM07] is manifested as:

> *In our opinion, it will not be possible to make very much progress in the analysis and understanding of these complex [written] narratives until the simplest and most fundamental narrative structures are analyzed in direct connection with their originating functions. We suggest that such fundamental structures are to be found in oral versions of personal experiences: not the products of expert storytellers that have been retold many times, but the original production of a representative sample of the population.*
>
> LABOV AND WALETZKY [LW97]

Thus, Labov's theory of narrative analysis defined the three elements of the narrative structure: the orientation, the complicating action, and the evaluation. More recently, this structure has been refined by Labov [Lab06; Lab72] to include additional elements: the abstract, the resolution, and the coda. Prior to Labov's theory, Freytag [Fre94] had proposed a dramatic structure containing five categories: Exposition, Rising Action, Climax, Falling Action, and Denouement. Several such theories with different structural labels were proposed [LW97; Lab72; Lab06; Pri12]. However, certain elements of the narrative structure are correlated across narrative theories. For example, Bruner's "breach in canonicity" [Bru91b] could correspond to (a) Freytag's "climax" – referring to the "turning point" of the fortunes of the protagonist [AH14] or (b) Labov's "most reportable event" (MRE)– describing the

event that has the greatest effect upon the goals, motivations and emotions of the characters (participants) in the narrative [LW97; Lab06] In fact, Labov argued that the entire story's purpose is to serve the MRE. Similarly, both Freytag's and Labov's theory of narrative structure contain "resolution" as one of the elements referring to the event that leads to a swift drop in the tension created by the MRE as one of the final aspects in the narrative. Therefore, it is possible to identify a functional schema that reconciles with multiple theories capturing both dramatic tension and the social aspect of online personal narratives.

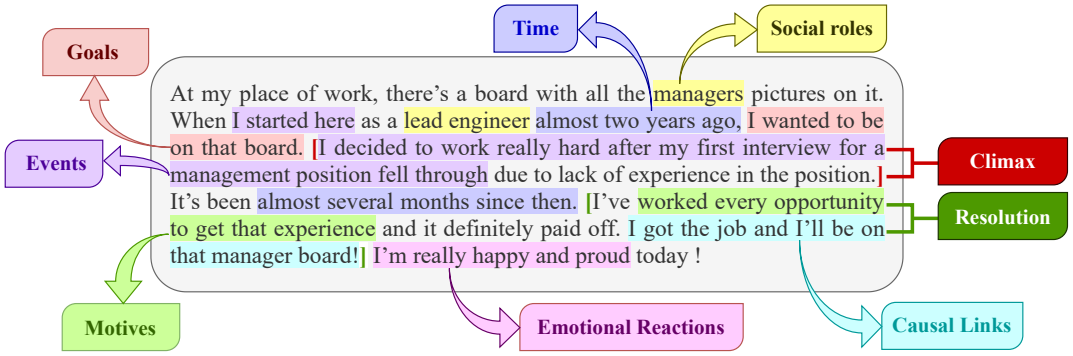


FIGURE 1.2: Sample Personal Narrative containing highlights of some of its major characteristics – events with actions, temporal order (time information), causal links, social roles, protagonist's goals, motives & emotional reactions, to list a few.

Figure 1.2 shows a tagged sample narrative from Reddit. This excerpt sheds light on some of the major characteristics of personal stories, given as follows:

- **Events**: Narrative contains a sequence of events with actions forming the outline of the narrative. We give examples of few related events in the narrative – "*I started here as a lead engineer*" and "*I worked really hard...*", "*my first interview for management position fell through*".

- **Temporal Information**: It refers to the timeline of events, most often presented in a piece of chronological order information (may also mention "time" details). For example, The following sequences provide time

information and order in which they occur – "I started here as a lead engineer *almost two years ago*", "I decided to work really hard *after my first interview..*", and "It's been *almost several months since then*. I've worked very hard...".

- **Causal Links**: Story events are generally united using causal connections. A causal link in our example narrative where one event could trigger the next event in the story is – "*I got the job*" $\mapsto$ "*I'll be on that manager board*".

- **Social roles**: The behavior and psychological states of a character in the narrative are determined by capability, situation, and social role. In the example narrative, social role information like "lead engineer" or "manager" plays a role in shaping their goals, motives, and emotions behind their actions.

- **Mental States**: This indicates the mental states (goals, motives, emotional reactions, desires, beliefs, etc.) of the characters in the narrative. The actions of a character could potentially cause mental state shifts depending on the social situation given in the event. Some of this may be explicitly expressed in the narrative, while others require inference abilities to identify inherent mental states. Narrator/Protagonist's mental state in the example narrative is given as follows – the phrase "I'm really *happy and proud*" represents the emotional reaction, "Goal state" is explicitly expressed as – "*I want to be on that board*" (implicit inference: to become a manager), "Motive" behind an action is – "worked every opportunity *to get the experience*". Here, the goal state refers to the cognitive representation of a desired end state [FF07]. Underlying all of these goals, though, is motivation, or the psychological driving force that enables action in the pursuit of that goal [AGL35].

- **Narrative Structure**: Correlated across prior narrative theories, we point out to the key elements of narrative structure – MRE/climax and resolution. In Figure 1.2, we mark the boundaries of climax and resolution with red and green square braces ([...], [...]) respectively.

In the past, different computation models leveraging the characteristics of narrative data have been developed towards enhancing aspects of story understanding capabilities[Bam12; Fic+17]. Some of these works include modeling event schemata and narrative chains, narrative causality, characters' interpersonal relationships, narrative and open domain question answering, understanding narrative structures, learning scripts, story plot generation and creative or artistic storytelling, story ending prediction, to list a few [VVZO15; SCM16; Joc13; Fin12; CJ08]. Several such approaches have emphasized the importance of (a) modeling the effect of events on characters' mental states on one hand [TFS95; Kal+02] and (b) interpreting causal and temporal relationships between events and identifying the ensuing structures in terms of the characters' motivation and affective states, on the other [Ger13; RTC05; Hap94]. Of particular interest in this dissertation, we explore these aspects by relying on personal stories collected from social media.

1.4 Research Problems

As described in Section 1.2, the focus of this dissertation is to develop and evaluate computational functional modules that exploits the link between mentalizing and narrative comprehension to capture specific aspects of socio-cognitive processes. Towards our goal, we formulate the research problems and propose solutions to overcome a subset of their associated challenges by applying information extraction, data mining, and deep learning techniques on an open-domain dataset of personal narratives from social media.

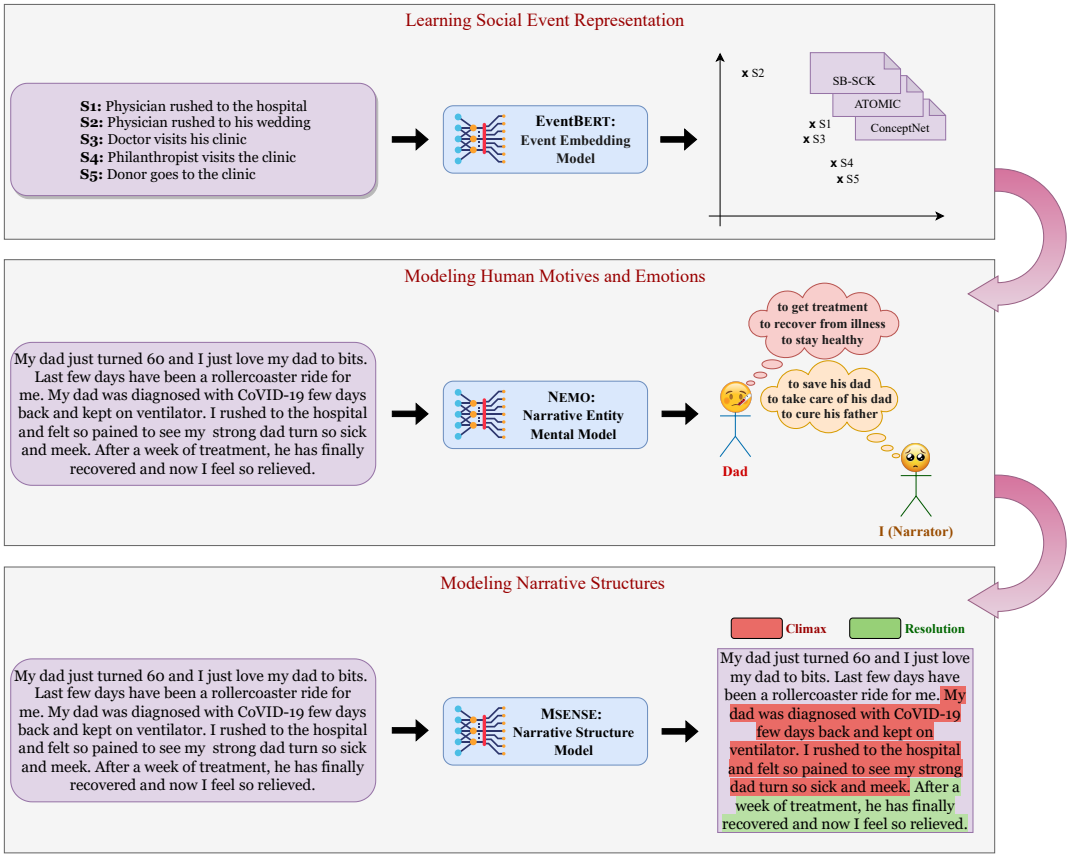


FIGURE 1.3: Overview of the Research Tasks tackled in this dissertation. (a) Social Event Embeddings: Sample event text inputs are provided on the left, and the expected relative positions of event texts in the learned embedding space are shown on the right. (b) Narrative Entity Mental Model: Example narrative text input is shown on the left, and the sample generations of intents of dad and son (narrator) are given on the right. (c) Narrative Structure Model: On the left, we display the input narrative text. On the right, we have the output of the sentence labeling task highlighting the climax and resolution sentences.

Figure 1.3 gives an overview of the research problems we aim to tackle in this dissertation. We present a brief outline of the research tasks (RT) and the proposed methods as follows:

- **RT1: Learning Knowledge-Enriched Social Event Representation**

  Personal narratives consist of an account of a series of causally-related events or experiences from the narrator's point of view. These experiences usually unfold naturally into a temporally extended daily event. The meaning of an event sentence or phrase can vary depending on

several factors like the speaker's perspective or the related domain. Prior approaches have relied on syntactic and semantic features to learn distributed representation of structured events [LDL18a; LDL18b; GWC16; Mod16]. However, there are several shortcomings with these embeddings as they cannot efficiently capture the relationship between events that are closer at the pragmatic space beyond understanding within lexical, syntactic, or semantic representation space. Contemporary definitions of pragmatic aspects of language include behavior that includes social, emotional, and communicative aspects of language [Ada+05; Par+17]. In the context of event representations, the pragmatic properties can specifically refer to the human's inferred implicit understanding of event actors' intents, beliefs, and feelings or reactions [Woo76; HN78].

In order to incorporate pragmatic implications of events, we focus primarily on social events, i.e., events depicting social situations and interactions. This is pertinent for our larger goal of modeling personal narratives. For example, as shown in Figure 1.3a, we ideally expect the following two events to be similar in the embedding space: "*Physician rushed to the hospital*" and "*Doctor visits his clinic*". However, most of the prior embedding approaches put the following two events closer, though they are unconnected: "*Physician rushed to the hospital*" and "*Physician rushed to his wedding*". Though they have strong lexical overlap and contain similar action verbs, they refer to different events with different intentionality. Similarly, there is a difference in how humans process the following two event texts in Figure 1.3a: "*Doctor visits his clinic*" and "*Philanthropist visits the clinic*". While the former has an intent of attending to his patients, the philanthropist's action may involve an act of goodwill. The social role information (e.g., donor vs. doctor) provides additional information about how we understand the events.

Thus, it becomes important to understand each event through a pragmatic lens for better narrative comprehension, and improved mentalizing capabilities [PM16; BMJ17]. Therefore, we solve this problem of computing a social event embedding by extracting semantic properties from the event texts and integrating salient knowledge that encompasses the implicit pragmatic abilities. This is considered a precursor to developing mental models of characters in narratives.

To this end, we propose to train our models with social commonsense knowledge about events focusing specifically on the intents and emotional reactions of event participants. We obtain commonsense knowledge assertions from ConceptNet [SCH17], ATOMIC [Sap+19a; Ras+18a] and also by aggregating more noisy commonsense knowledge using web-based data mining techniques. This new commonsense knowledge dataset is referred to as Search-based Social Commonsense Knowledge (SB-SCK). Subsequently, we employ a fine-tuned BERT-based encoder, called EVENTBERT, to effectively embed social events with semantic and pragmatic attributes. Empirically, we demonstrate the capabilities of our social event embeddings by a strong performance on many downstream tasks like event similarity, reasoning, and paraphrase detection.

- **RT2: Modeling Human Motives and Emotions from Short Personal Narratives**

  One of the central objectives of this dissertation is to foster research in imparting mentalizing abilities to machines. With such a grand challenge, we take a small step in this direction by developing models that can embed and explain human motives and emotional reactions. Understanding a story not only requires keeping track of the sequence of events happening in the story but also inferring and interpreting the

mental states of characters and interactions between them. It is thus natural to consider the usage of stories towards building a model for inferring aspects of people's mental states from their actions in social situations.

One of the key challenges lies with difficulty to acquire annotated data containing explanations of stated or implied intents or reactions of characters to events in the narrative. Therefore, we address this challenge by adopting a combination of web-based data mining and information extraction (IE) strategies to automatically aggregate noisy expressions of motivations and emotional reactions related to specific events and social roles in the text. This results in a weakly-annotated dataset containing characters motivations and emotions from personal narratives. We also utilize the sentence-level social commonsense knowledge assertions, SB-SCK, already extracted for learning social event embeddings.

Next, we focus on devising models to continuously track shifts in mental states of character and effectively embed and generate their text explanations. Note that these explanations are not always stated in the narrative text, but the implied mental state explanations are derived using a knowledge enrichment and entity tracking module implemented using external memory. In Figure 1.3b, we show a sample narrative with two characters – a father and son (narrator). The son (narrator) describes his experience of his father getting diagnosed and recovering from CoVID-19. Take the sentence – "I rushed to the hospital", the intent of the narrator ("I") is "to take care of his dad" or "to be worried for his dad". To produce such explanations, it is necessary to consider the story context and social role because modifying the context and social role information could significantly alter the corresponding intent

and emotional reaction behind the same action (as shown in Figure 1.3a and described under RT1).

For our modeling purpose, we implement a Transformer-based encoder-decoder architecture, referred to as NEMO [1]. We augment this encoder-decoder architecture with our pretrained knowledge-enrichment module, EVENTBERT, to incorporate social commonsense knowledge. Finally, we employ dynamic state tracking of entities with the help of an entity-based memory module to produce contextual embeddings and explanations of characters mental states. Experimentally, we demonstrate the effectiveness of our model on a benchmark character psychology dataset called STORYCOMMONSENSE [Ras+18c] and a downstream empathetic response generation task.

- **RT3: Modeling Narrative Structure in Short Personal Narratives**
  Since personal stories obtained from social media are idiosyncratic and noisy, it becomes increasingly difficult to automatically interpret these stories. Moreover, they hardly adhere to any structural conventions in narratives. Following narrative theories proposed by Freytag, Prince, Bruner, Labov & Waletzky [Cut16; Bru91b; LW97], we draw on the prominent elements of narrative structure (MRE/climax and resolution) that correlate across multiple theories and are deemed appropriate for the informal nature of narratives. This work aims to leverage computational methods at the intersection of information retrieval, NLP, and aspects of psychology and automatically predict the key elements of narrative structure – MRE and resolution.

  We construct a STORIES dataset [2] corpus comprising personal narratives from Reddit with fine-grained manual highlights of climax and resolution. Drawing on prior studies regarding the relevance of the

[1] Short for **N**arrative **E**ntity **M**ental m**O**del
[2] Short for **ST**ructures **O**f **R**edd**I**t P**E**sonal **S**tories

protagonist's mental states, we hypothesize that mental state embeddings derived from our NEMO model could potentially provide an advantage in determining the boundaries of structural components of the narrative. Therefore, we train an end-to-end neural network called M-SENSE[3] as a sentence labeling task and infuse protagonist's mental state embeddings in addition to linguistic features though a multi-fusion feature technique to automatically identify these elements of narrative structure. In Figure 1.3c, we show sample prediction of climax (red) and resolution (green) by our M-SENSE model that exploits the narrator's (protagonist) shift in intent and emotional reactions for the classification task. Overall, we are able to achieve $\sim$ 20% higher success in detecting climax and resolution in short personal narratives than the previous state-of-the-art methods.

1.5 Overview of Contributions

The goal is to establish that endowing AI agents with social awareness can result in better performance across various tasks and environments and lead to the development of models that are better able to meet human preferences and adapt to new circumstances. In service of this goal, this dissertation makes the following contributions:

- A representation learning framework that learns to:
  - Disentangle semantic and pragmatic attributes from social event descriptions. Here pragmatic properties refer to the human's inferred implicit understanding of event actors' intents, beliefs, and reactions.

[3]Short for **M**ental **S**tate **E**nriched **N**arrative **S**tructure mod**El**

  - Encapsulate both the attributes into an embedding that can move beyond simple linguistic structures and dispel apparent ambiguities in the true sense of their context and meaning.

- Transferable social event embeddings demonstrating impressive performance gains in downstream tasks – event similarity, reasoning about social interactions, and paraphrase detection task.

- Automatic data acquisition method based on web-based data mining and information extraction (IE) strategies for constructing a corpus containing weak-annotations of implicit and explicit motivation and emotion text expressions. This can be applied to different textual domains.

- Two datasets – personal narratives corpus and a search-based social commonsense knowledge dataset (SB-SCK) for modeling human motives and emotions from short first-person stories. While the personal narratives corpus contains explicit characters' motivations and emotions conditioned on the entire story context, our SB-SCK comprises phrase or sentence-level implicit mental state mappings associated with social events.

- Transformer-based story understanding model augmented with external memory modules that infuse social commonsense knowledge and dynamically track entities' mental states.

- STORIES dataset consisting of short personal narratives with manual annotations of key categories of high-level narrative structure, specifically climax and resolution.

- An end-to-end computational model based on a multi-feature fusion approach for automatically identifying climax and prediction using the protagonist's mental representations.

The models The above research has also led to the production of research artifacts, including peer-reviewed publications [VR21a; VR21b], datasets and tools to support these projects [4] and make them reproducible for other researchers.

1.6 Organization of Dissertation

The rest of this dissertation is organized as follows:

- Chapter 2 provides details of pertinent background information and a survey of the related work. First, we explore the literature that explains the mutual relationship between narrative processing and ToM and how they play a critical role in understanding human social behavior and promoting social cognition. Next, we discuss the recent works on representation learning for social events and examine other relevant studies that have benefited from commonsense knowledge. Following that, we summarize other story understanding and text generation efforts that are directed towards inferring characters' psychological states. Finally, we refer to prior approaches in the field of narratology and NLP that investigate the story structure based on narrative theories and describe the techniques used to eventually predict the core components of narrative structure.

- Chapter 3 presents the datasets constructed as a part of this dissertation. We describe the data collection pipeline and the annotation setup employed to obtain data for training and evaluating our models. For

[4] https://github.com/pralav

modeling social events and characters' mental states, we delineate web-based knowledge mining and information extraction techniques to automatically aggregate weakly-annotated data. Next, we outline our annotation setup that allows for fine-grained span-level highlighting of key elements of narrative structure, i.e., climax and resolution in short personal narratives. We explain different span and sentence-level measures to compute inter-annotator agreement for our manual data collection task and show that we can achieve a substantial inter-annotator agreement for both the categories of narrative structure.

- Chapter 4 proposes a representation learning framework that effectively embeds both semantic and pragmatic aspects of social events with the help of a growing set of social commonsense knowledge assertions acquired from different domains. We utilize knowledge assertions from ATOMIC [SAP+19A], CONCEPTNET [LS04] and our aggregated dataset, SB-SCK [VR21a], for training our BERT-based social event embedding models. Experimentally, we demonstrate the benefits of applying our social event representation in various downstream tasks like event similarity, reasoning, and paraphrase detection tasks.

- Chapter 5 describes a Transformer-based architecture to model characters' motives and emotions from personal narratives. We develop a model that learns to produce contextual embeddings and explanations of characters' mental states by integrating external knowledge along with prior narrative context and mental state encodings. We leverage the social event embeddings explained in Chapter 4 and utilize the weakly-annotated personal narratives to train our model and demonstrate its effectiveness on the benchmark character psychology dataset. Additionally, we show that the learned mental state embeddings can be applied in downstream tasks like empathetic response generation.

- Chapter 6 analyzes the importance of studying the links between cognitive and linguistic aspects in narrative comprehension. In this chapter, we probe this interdependence by jointly modeling textual semantics and mental language in narratives for improved detection of narrative structure categories. We implement an end-to-end computational model that leverages the protagonist's mental state information and integrates their representations with contextual semantic embeddings using a multi-feature fusion approach to model high-level narrative structure. We show that our model surpasses several prior zero-shot and supervised baselines for identifying climax and resolution.

- In Chapter 7, we first summarize the contributions of this dissertation. Next, we highlight the limitations of our work and discuss the directions for future research.

Chapter 2

Background & Related Work

Our current research explores how stories are central to human cognition and how they influence our understanding of the world and events that unfold around us. The notion that narratives could facilitate inference of others' mental states, referred to as mentalizing has very early origins[Hak00]. As early as 330 BCE, Aristotle mentions in Poetics that "man tends most towards representation and learns his first lessons through representation". Poetics is one of the earliest surviving works of dramatic theory, elucidating it as a language that represents and imitates life. Further, he argued that stories convey reality about the world even though fictional stories may not be a completely accurate representation of truth [Oat99]. In the field of psychology and certain theories in narratology, narratives are described as representations of temporally coherent events pivoted around the goals of a protagonist, typically following a schema or structure consisting of elements, including a setting, an inciting incident, a rising action, a resolution, and a denouement [TVDB85; Rum75].

Since stories are commonly about people, their mental states, and their relationships [CS96], mentalizing or Theory of Mind might be one of the social cognitive processes engaged by narratives [Hog03; Zun06] involving the ability to infer beliefs, thoughts, motives, emotional reactions of other people. In this chapter, we focus on reviewing the previous studies that aspects of narrative comprehension, mentalizing, and the well-established positive

impact of their relationship.

In the following sections, we discuss several interdisciplinary studies that present the importance of mentalizing in navigating the social world and its influence on narrative comprehension. Following this, we survey computational approaches investigated earlier on each of these fronts and provide a lead to how our work tackles some of the existing challenges towards accomplishing our overall goal.

2.1 Background

Overview of Mentalizing and Its Relevance to Narratives

The ability to anticipate, represent and reason about what others will think, feel or do in different situations is central to social cognition. Consider a scenario where one experiences difficulty predicting social signals or implications like agreeable people tend to be courteous and warm or exhausted people tend to show anger; it can lead to a complicated social life filled with misconceptions, faux pas, and miscommunication. Fortunately, humans can predict others' probable social actions through either their personality traits (e.g., agreeableness) or mental states (e.g., tiredness). Neuroimaging studies have also suggested the mentalizing, or "theory of mind" network plays a role in social cognitive processing more broadly, including reflecting on personality characteristics of one's self and others, inferring mental states including emotion processing and intentions from actions.

A model proposed by Shamay-Tsoory et al. [STAP07] divide ToM into two separate systems, namely cognitive ToM and affective ToM. Cognitive ToM is described as involved in processing inferences about others' beliefs and intentions, whereas affective ToM is involved in processing inferences about other peoples emotions and feelings. This model describes affective

and cognitive ToM involving common and different brain areas studied by Poletti, Enrici, and Adenzato. Several studies in the field of personality psychology [Ryc04] have some congruence with the above idea of cognitive-behavioral models. However, in these studies, personality is defined as "A dynamic and organized set of characteristics possessed by a person that uniquely influences their cognition, motivations, and behaviors in various situations" [Ryc04]. The dimensional theories like the "Big Five" model or theories that propose additional six personality dimensions are known to be tailored to understand stereotypes, mind perception, and common behaviors. A work by Tamir et al. [TT18] studied if these low dimensional personality dimensional theories can efficiently aid social predictions. It was found that much of the richness of others' minds can indeed be compressed to coordinates in low-dimensional trait space. Similarly, there is a reasonable amount of literature [Tam+16a; GGW07] that support how the representation of people's momentary mental states into lower dimension can facilitate social prediction in humans.

It is often highlighted how engagement with narratives encompass a deeply embodied mental simulation [Zwa04] and how narratives offer encapsulated abstract representations of concrete work scenarios and people. Several studies point to a common implication that stories help to foster a better understanding of other people. The relationship between narratives and social cognition has been investigated with school children [Mar+06]. Given the expectation that children who are exposed to more stories tend to develop mentalizing capabilities more rapidly than other children, different approaches have been used to study this hypothesis, and the results have turned in support of this notion. Maternal expertise in choosing children's literature predicted better empathy, socio-emotional adjustment, and improved false-belief reasoning in children [MTM10; AA09].

Kimhi [Kim14] discussed the development of mentalizing ability across

the life span in persons focusing on its social and academic manifestations that are critical for everyday life skills. Considering the social manifestations of ToM in symbolic play, conversation, and autobiographical memory and academic manifestations of ToM in reading comprehensions, narrative skills, and writing abilities, many related works, including the literature with mixed evidence on the significance [BCEGB00] and direction of association [JA00; Suw+12] were discussed. However, there is consensus that children's ability to understand others mental states, though it may not be sufficient by itself, appears necessary for engaging in adaptive and positive behavior [Ast04] and these abilities are reflected in their social interactions [Hug98]. This is further supported by results from an assessment of individuals with autism. Even their high verbal and intellectual levels do not aid in navigating effectively in social and academic settings exacerbated by the diminished attention to social cues and difficulty in social adaptive behavior [Kli+02; BZ13].

Consequently, interventions have been proposed and developed to enhance ToM in children and young people with autism spectrum conditions. Specific ToM socio-cognitive training [Gou+11; PP13] (e.g., Thought Bubble Training) has been found to enhance the targeted skills; yet, generalization to other skills generalization to the natural environment has been minimal for the most part. More sophisticated interventions (e.g., dyadic & group social interventions) involve training strategies [MKD07; Bau07; BZ+13] that integrate social interaction training in children's natural settings with the main social interactive agents (teachers, peers, and parents) involved along with specific sociocognitive abilities. With such training and improvement in social cognition, language, and self-regulation [Len03], it was observed there was a general decrease in children's aggressive behaviors and an increase in pro-social acts throughout their preschool years [FS73; Per+07].

Construed broadly, mentalizing covers a range of capabilities such as

perspective-taking, simulating mental states, identifying character traits, social and emotional reasoning. Linked closely to the acquisition of social vocabulary and processing of social information, mentalizing is critical for facilitating active engagement in social and academic activities. Taking a leaf out of the various experiments explained above, we are interested in developing learning techniques directed at analyzing stories to augment mentalizing skills. Therefore, in this work, we incrementally aggregate social commonsense knowledge in the form of intents and emotional reactions and further improve the narrative processing skills. For our dissertation, we use the aggregated knowledge and also adapt quickly to the different social contexts by harnessing the transferability to personal narratives. We, therefore, focus on addressing a subset of challenges to represent people's mental states or personalities and draw insights into people's social behavior.

2.2 Modeling Human Social Behaviour

In the light of our work, it is crucial to review the literature that adopts different mechanisms to model human social behavior. We investigate the background work on a set of methods in the context of building mentalizing abilities towards a long-term goal of socially intelligent systems. With increasing human-machine hybrid technologies, the real-world interactions with AI systems are often stilted. It is essential to acknowledge the challenges associated with the understanding of explicitly unstated desires, emotional states, and intentions of users from language. Misinterpretation of users' implied intents and implicit beliefs from natural language could have dramatic real-world consequences. Building AI systems that can interact with humans fluently will require machines to share common knowledge about how people will act, communicate and react under specific contexts and circumstances.

Many AI researchers have attempted to adopt these ideas and build systems that can encode personality traits or mental states into representations and utilize them in different social contexts. Bridewell and Isaac (2011) [BI11] introduced a computational framework for common, complex, and under-investigated aspects of human social behavior like deception based on the capacity to reason about the goals of other agents, resting on mental state ascription. Fahlman [Fah11] proposed a knowledge-base system, Scone, used to emulate some aspects of human mental behavior and support human-like commonsense reasoning and language understanding. Beyond domain-specific knowledge, social understanding requires generic knowledge about social interactions and their ensuing effects on mental states. Early research conducted by Wilensky along these lines inferred the intentions of interacting agents while Dyer dealt with extracting morals from social scenarios. Winston's [Win14] Genesis system was developed to understand and generate stories using computational models that use commonsense inference rules and concept patterns. This includes their work to support question answering, personality or mood-based interpretation, and summarization of stories.

One possible direction explored to overcome shortcomings of AI systems in navigating the social world is to endow them with commonsense knowledge. While there have been significant efforts to create knowledge bases like Cyc [Len+90] and ConceptNet [LS04], there is a paucity of inferential knowledge related to people's behavior in the form of their motivations and their reactions. Using stories to define a space of acceptable behaviors, Harrison et al. [HR16] developed a technique to prevent autonomous agents from exhibiting anti-social or psychotic behaviors. Recently, knowledge bases such as Event2Mind [Ras+18a], and ATOMIC [Sap+19a] are tailored to capture the mental states of people linked to day-to-day events. Another line of work towards improving automatic recognition and interpretation of human social signals in AI systems relies on inferring personality traits. Considering that

personality compels a tendency on many aspects of human behavior, mental states, and affective reactions, there is an enormous opportunity for sensing spontaneous natural user behavior to facilitate efficient interaction in social settings. [SHR18] presents a hypothesis that users with similar personalities are expected to display mutual behavioral patterns when cooperating through social networks. Imitating personal style in dialogue systems has demonstrated promising results. Some of the early efforts include modeling personas of movie characters and incorporating speaker persona in dialogue models based on speaking style characterized by natural language sentences [BOS13; Li+16]. Despite recent successes, it is still incredibly challenging to build socially intelligent agents that can understand humans and engage in socially competent conversations that involve empathy, cooperation, persuasion, care-giving, to list a few.

Recent approaches have focused on reflecting upon the concepts of human ToM to attribute mental states such as intentions and beliefs to inanimate objects. Some notable approaches include those that use hierarchical Bayesian inference [Bak+17; YDF08; BST11] or artificial neural networks [LP18; Lan+17b]. The former is generally cognitively-inspired and suggests the existence of a "psychology engine" in cognitive agents to process ToM computations, while the latter achieves imparting ToM to a certain degree by characterizing different species of deep reinforcement learning agents. In addition to these methods, there also have been multi-agent models rooted in statistical machine learning theory and robotics. These approaches have generally evaluated simulations of the theory of mind in relatively simple situations. However, there is very limited work in this area of combining the theory of mind and language. It is also well known that two of the most fundamental elements of human cognitive capabilities are the ability to communicate through complex language systems and the attainment of a theory of mind. Interestingly, both language and theory of mind develop relatively

at the same time in a persons life. Language is a fundamental element in understanding emotions, thoughts, and actions that are constitutive of both experiences and perceptions. Early ToM abilities facilitate the development of early language abilities, while more complex language abilities are a precursor to complex mentalizing (or ToM) abilities. Language, ToM, and social skills are all connected and interdependent. Hence, we focus on bridging this gap in research towards understanding language and the development of mentalizing abilities.

Towards this goal, it is important to produce an efficient social event representation that can contribute to modeling the motives and emotions of characters. The primary reason behind this is because narratives consist of a sequence of events about the social situations presented to the characters in the narrative [AGS83; Ger13]. Understanding the meaning of social events requires representing them at syntactic, semantic, and pragmatic levels to embed them in the context of commonsense knowledge. An ample amount of studies has been centered around constructing situation models to understand a text and the events described in them [MM09; Zwa+98]. Situation models involve dynamic processes that allow them to fuse information from the event text description with world knowledge and produce an integrated event representation [Zwa+98; ST93]. Thus, such representations have to be extended to narratives by considering the story context, entities, and actions and how they are connected through events and situational relationships. Such approaches enhance AI systems' capabilities to better recognize characters' planned actions and their intentionality towards achieving desired states [TBS89].

Many computational approaches have been attempted to model the motives, emotions, desires, and goals of characters in the narrative. A recent body of work [Gui+17; Gho+17] is related to detecting emotional stimulation in narratives and utilizes specific attributes like sentiment or affect states

based on LIWC categories. A close work used to infer character mental states in short five-sentence commonsense stories based on rich low-level annotations of intent and emotions [Ras+18c]. In our work, we develop automatic techniques to extract weakly-annotated mental state expressions from natural variable-length personal narratives and propose a method to leverage aggregated social commonsense knowledge to efficiently generate explanations of motivation and emotion states of characters in the narrative. Further, we address the modeling challenges by incorporating social commonsense knowledge from social events and employing entity modeling for tracking the mental states of characters in the narrative.

In this dissertation, the first aspect of our research will focus on learning to model the mental states of people by integrating commonsense knowledge of social behavior with knowledge acquired from a textual narrative corpus. The representations learned by such models are more likely to yield AI systems that are generally better at perceiving, understanding, and responding effectively to different social situations. With commonsense knowledge acting as the basis of mentalizing, the behavior of such socially-aware systems, specifically during human-machine social interactions, will be consequently more recognizable and aligned to people's expectations.

2.3 Modeling Narrative Structure

Narrative theory has drawn distinctions between the story's content or theme, that is, the narrated event and its form, or telling [SW17]. This notion of conveying the same story in numerous forms is drawn from work in many academic traditions, including literary studies, folklore/anthropology, psychology, and sociolinguistics [McQ00; DF08]. For example, literary analysis has examined the structure, cultural forms, and textual qualities of narrative (often literary texts), while anthropological studies have explored the

content and form of stories, their cultural resonance, and the storytelling practices of different cultures [Pro10; Pol81]. Coherent situation models of a narrative require the ability to accurately recognize boundaries in narrative episodes which can be attributed to structural components of narratives [Ger13; Mag+12; MTK05]. Recognizing temporal shifts in mental states and monitoring them is critical towards understanding the boundaries between narrative episodes represented in the situation models [Zwa+98; Zwa16]. Drawing from narrative psychology, narratives have been used to understand cognitive processes. However, the protagonist's mental models can be an intriguing way of imposing structure in narratives [Bru90; Bru91b; Bru91a; Bru09].

Several previous works have laid emphasis on understanding different aspects of narratives [Els12a; GBS11; Fin12]. Earlier many annotation schemas such as Rhetorical Structure theory [MT88] and Penn Discourse Treebank [Pra+08] were proposed to analyze different types of discourse. Such schemata provide a principled way of performing structural analysis of text [BHN18; KG78]. Given the limited efforts towards capturing the functional schemata or structure in narratives, it is advantageous to undertake computational interpretation of such structures in order to comprehend the meaning conveyed in a narrative. These structures typically indicate a list of functions that follow a specific order sequence between them. Moreover, they could potentially signify critical points of the narrative text and contribute to the dramatic arc [OR11; Li+17].

Propp [Pro10] defined the repeated plot elements as functions for Russian folklore. These are understood as a part of characters' acts, defined from their significance for the course of the action. There are other theories like Campbell's "Hero's journey". The common property of such theories is they are closely associated with particular kinds of stories and domains. Lehnert [Leh81] presented plot units as conceptual structures for modeling events

Freytag	Labov & Waltezky	Prince
Exposition	Orientation	Starting State
Rising Action	Complicating Actions	
Climax	Most Reportable Event	State-Changing Event
Falling Action	Resolution	Ending State
Denouement	Coda	

TABLE 2.1: Correspondence between categories from different narrative theories as in [Li+17]

and states in the stories and their corresponding relationships between them. The primary motivation behind the idea of plot units is the notion that emotional reactions are central to the narrative, and the story's plot can be tracked using the transition of affect states at the event-level. However, this assumption has its limitations where affect states emerge from the events, and mental states are not modeled or distinguished from the actual events occurring in the story. To overcome the limitations of the plot units, Elson [Els12a] presented a richer annotation schema, referred to as Story Intention Graph (SIG), to capture timelines as well as beliefs, intentions, and plans of story characters. It consists of three layers and is highly expressive, involving motivation and affect states of characters. However, the level of expressiveness expected from this approach is highly resource-intensive and is sometimes difficult to interpret and annotate.

Other narrative theories generalize stories across genres to contain a certain uniform structure. Prince [Pri12] proposed three basic states which describe the narratives to contain a beginning, a middle, and an end. Here the middle acts as the transformational event. Similarly, Freytag's dramatic pyramid contained five parts that include – Exposition, Rising Action, Climax, Falling Action, and Denouement [Fre94]. Similarly, Labov and Waletzky [LW97] proposed a theory on oral narratives which initially divided narrative clauses into three dimensions – temporal, structural and evaluation

points in narratives [LW97; Lab06; Lab01]. Here, the complicating action culminates in the "Most Reportable Event" (MRE), which indicates the event with the highest cognitive tension that the characters grapple with. The structural label Evaluation is claimed to end with Resolution and Coda. Based on [Li+17], correspondence across narrative theories was identified, and these categories and their related counterparts in other theories are shown in Table 2.1. Rahimtoroghi et al. [Rah+14], and Swanson et al. [Swa+14] used a subset of Labov's categories, including orientation, action, and evaluation in personal weblog narratives. Black and Wilensky (1979) evaluate the functionality of story grammars in story understanding, [PKL19] introduced a dataset consisting of screenplays and Wikipedia plot synopses annotated with turning point as a means of analyzing their narrative structure. More recent work by [Lev+20] addressed the task of automatically detecting narrative structures primarily directed at news stories. By adopting elements from the narrative theory of Labov and Waletzky (Complication and Resolution) and designing their new element, they construct a news corpus and proposed supervised methods to identify them.

Our goal is not directed towards building a new functional schema for social media personal narratives in this work. However, our primary objective is to test the hypothesis that the mental state representation models can significantly impact improving narrative comprehension tasks, which in our case, is identifying key elements of narrative structure in short personal narratives obtained from social media. Given the different theories, their labels, and commonalities, we prioritize climax/MRE and resolution to be the categories of interest for our work. The intuition behind selecting these two elements lies in the aspect of 'tellability'. Researchers in narratology have analyzed various components of a narrative that contribute to a notion of plot quality referred to as 'tellability'. It is commonly derived from certain structural properties used in narrative theory. Bruner insisted on the fact that "to

be worth telling, a tale must be about how an implicit canonical script has been breached, violated, or deviated from". Bruner's 'breach in canonicity' [Bru91b] could correspond to (a) Freytags 'climax' – referring to the 'turning point' of the fortunes of the protagonist [AH14] or (b) Labov's 'most reportable event' – describing the event that has the greatest effect upon the goals, motivations and emotions of the characters (participants) in the narrative [LW97; Lab06]. Moreover, most of the narratives containing an event of highest tension also reach a 'resolution' stage involving a swift drop in tension as the final step. Our work aims to develop computational approaches that model the key elements of narrative structure – MRE and resolution. Drawing ideas from prior theories that express the influence of protagonist cognitive state [Els12a; Els12b; Bru91b; Bru09], we rely on the fine-grained mental states of the protagonist in the narrative and compute the shifts in their inner states over time for identifying key narrative events and boundaries that effectively contribute towards the automatic prediction of different structural components of the narrative.

Chapter 3

Datasets & Annotations

Several previous studies [Els12b; Oat95; CJ08; PM14; CGDI16] on narratives have used different forms of textual data ranging from news stories to literary texts to Wikipedia articles as described in Chapter 2. Given our overarching goal of investigating the interplay between narratives and mentalizing, we pivot our work to take advantage of the characteristics of personal narratives as discussed in Chapter 1. In this chapter, we discuss the various datasets collected for the purpose of our research and describe in detail the strategies used to extract weak-annotations of relevant information from the data. By aggregating personal narratives from Reddit, we process these narratives to derive specific properties from them, essential for our modeling purposes.

We prepare three main in-house datasets as a part of this work namely Personal Narratives Corpus, Search-based Social Commonsense Knowledge (SB-SCK) dataset, and STORIES[1] corpus. These datasets are central to our modeling and evaluation phases. In addition to these datasets, each research problem we proposed in Chapter 1 utilizes other publicly available benchmark datasets for evaluation. We describe our data collection process in two parts.

- We aggregate explicit and implicit expressions of motives and emotions at the sentence-level with and without the story context. Implicit intents and emotions are obtained using web-based mining without

[1]Short for **ST**ructures **O**f **R**edd**I**t P**E**sonal **S**tories

Datasets	Annotation Type	Size	Dataset Details
SB-SCK Dataset	Automatic	$\sim 100,000$	Sentence-level implicit mental state knowledge mappings.
Personal Narratives Corpus	Automatic	$\sim 85,000$	First-person Reddit stories with weak-annotations of explicit motivation and emotion expressions.
STORIES Corpus	Manual	$\sim 2,500$	First-person Reddit stories annotated with Climax and Resolution.

TABLE 3.1: Summary of the consolidated in-house datasets used in this dissertation.

any story context, and these are applied for embedding social events (SB-SCK). Explicitly stated expressions of intent and emotions are usually extracted along with the story context to model the characters' mental states (Personal Narratives Corpus).

- We construct the STORIES corpus to identify critical elements of narrative structure in short personal narratives. Since climax and resolution are predominantly present in most stories, we let the crowdworkers to manually select portions of the story that qualify as climax and resolution resulting in a dataset containing fine-grained manual annotations.

Table 3.1 provides a summary of the in-house datasets aggregated, processed, and partially annotated for our use in this dissertation. In the following sections, we delve deeper into the data collection processes drawing ideas from information extraction and data mining techniques.

3.1 Extracting Motives and Emotional Reactions

Our data collection pipeline is depicted in Figure 3.1. We aggregate two datasets: (a) weakly-annotated personal narratives corpus and (b) Search-based Social Commonsense Knowledge (SB-SCK). The former is intended to capture the motives and emotions extraction considering the entire story context. These are generally explicitly mentioned by the narrator in their stories. One of the limitations of the personal narratives corpus is that it may not contain implicit mental state mappings (motives & emotions) for several events in the narratives. To alleviate this limitation, we collect sentence-level implicit mental states by adopting a combination of web data mining and information extraction strategies. We elaborate on the steps involved in our data collection process in the following sections.

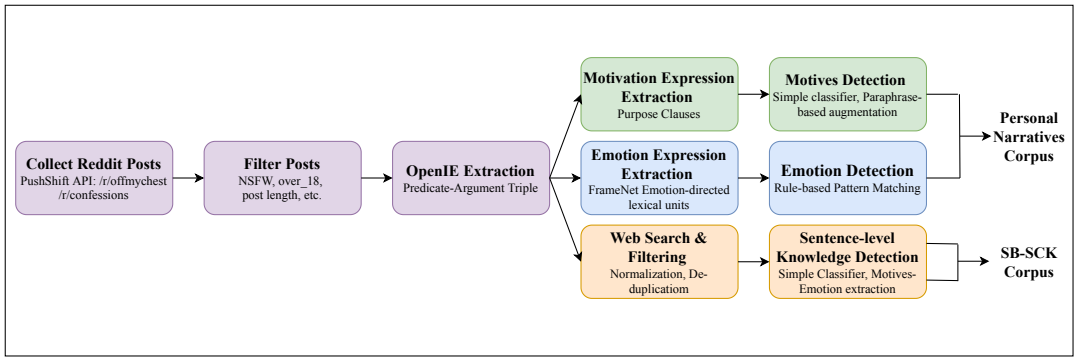


FIGURE 3.1: Illustration of data collection pipeline.

3.1.1 Personal Narratives Corpus

We construct a corpus of personal narratives by gathering posts from Reddit related to daily interactions, life experiences, relationships, comical or embarrassing situations, to name a few. Using Pushshift API[2], we aggregate 887,441 posts from specific subreddits: /r/offmychest and /r/confessions. Of these posts, we discard all those posts with tags like "[Deleted]", "NSFW"

[2]https://pushshift.io/

[3] or "over_18" field set to true. The number of sentences in the posts ranges from 1 to 1015. Further, we remove texts containing less than three sentences, based on Prince's definition [Pri12] of a minimal story as consisting of a starting state, an event, and an ending state. We compute the 90^{th}-percentile of the story lengths and remove those that exceed this length. This augurs well for our specific interest in short personal narratives. Therefore, we are left with 439,408 posts, with an average length of 12.08 sentences.Figure 3.2 shows the data distribution related to their lengths.

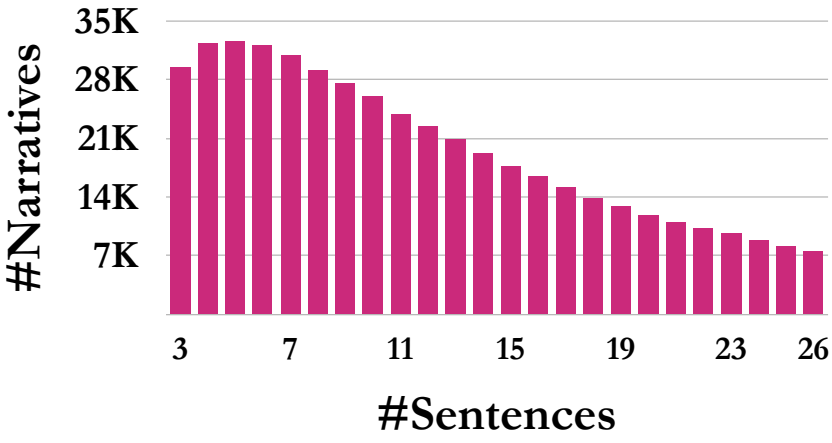


FIGURE 3.2: Dataset details: Personal Narrative Statistics – No. of narratives w.r.t their lengths.

To create our dataset related to motivations, we look for specific expressions associated with intents or purpose. Human motivations and emotions can be expressed linguistically in many ways, sometimes with explicit use of purpose clauses. Generally, purpose clauses take the form: To-Infinitive; (In order/So as) + To-Infinitive, (so that) + Subject + Verb; For + Noun/'ing'-form. In order to systematically identify text expressions that specify motivation, we leverage OpenIE[4] methods [Sta+18; APM15] to extract a list of propositions usually composed of a single predicate and an arbitrary number of arguments. Using PropBank [PGK05] and its annotation scheme, we can

[3] NSFW – not safe for work
[4] https://demo.allennlp.org/open-information-extraction/

break down syntactically complex sentences as: (a) ARG-0 related to the argument exhibiting features of prototypical agent and (b) ARGM-PRP related to the purpose or motivation expressions in the text. Figure 3.3 shows a sample OpenIE extraction of agent and its purpose/emotion.

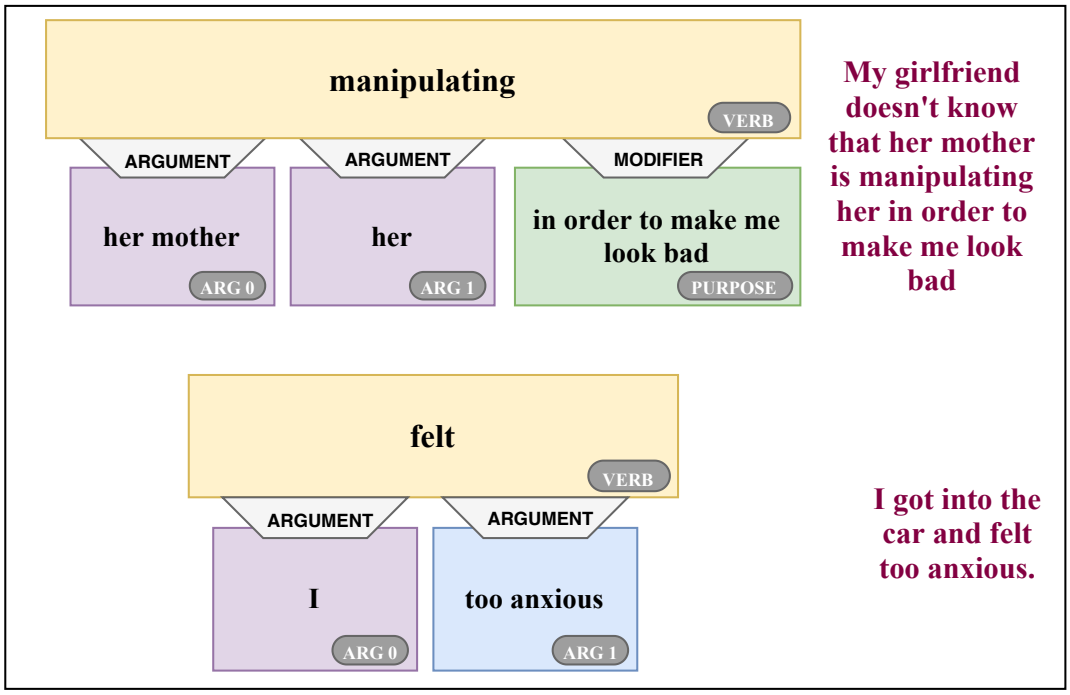


FIGURE 3.3: Sample OpenIE extraction containing arguments referring to agent and their motivation (purpose) and emotion.

One of the authors assessed the extraction quality by analyzing a random subset of the agent-purpose pairs for each type of purpose clause and their context. We manually identify a set of 300 extracted purpose clause texts if they genuinely reflect the motivation behind an action. To filter trivial motivation expressions (e.g.,"to do it"), a logistic regression classifier is trained by constructing hand-crafted features from text like mean word embeddings, POS tags, number of words, presence of stopwords, and entities. Eventually, we shortlist those expressions above a threshold score, $\rho_{pn} >= 0.4$. By eliminating trivial extractions with a basic classifier (see Appendix **??**), we use

the filtered data as our weakly-annotated training data. Further, we augment these extracted motivation texts with their paraphrases using a back-translation approach [Edu+18] to simulate multiple-annotation settings. A pretrained English↔German translation model is used for this purpose (e.g., to divert attention in tough times → to distract attention in difficult times).

We adopt a similar strategy to extract the emotions of characters in the narratives. First we identify 400 keywords extracted from a combination of: (a) emotion-directed[5] lexical units from FrameNet [BFL98] corresponding to different emotions, and (b) emotion vocabulary list[6]. Though we don't have any semantic role labeling for emotions, we still feed the sentences through the OpenIE extraction method. By examining extracted propositions, we discard those story sentences when: (a) sentence is negative (contains not), (b) emotion keyword is not a part of the predicate, and (c) the first argument is neither a noun nor pronoun. Using the first argument as the agent experiencing the emotion and lexical units specified in FrameNet to express feelings footnoteWe choose semantic frames related to "Feeling" (e.g., verbs like feel, experience, get, be; phrases like sense of, feelings of, full of), we map specific sentences in the narrative to the particular character and its emotion expressions. We accomplish this by utilizing spaCy's rule-based matching tool[7] to capture particular patterns in text. The data statistics are given in Table 3.2. Sample extractions are highlighted in Figure 3.4.

Three non-author annotators labeled a random sample of 300 instances (balanced between intent & emotions) for validation. Given the narrative context up to the sentence of interest, each annotator is asked to choose the right intent or emotion explanation expressed or implicitly felt by the character in the narrative. We let the annotators choose from the candidate texts that are: (a) extracted using our method, (b) chosen randomly, or (c) None (if

[5]https://framenet.icsi.berkeley.edu/fndrupal/luIndex
[6]https://www.enchantedlearning.com/wordlist/emotions. shtml#wls-id-0
[7]https://spacy.io/usage/rule-based-matching

Personal Narratives	
Intents	I haven't been able to get my degree, and it's killing me. I don't know where to start or how to do it. I put my job in stand by in order to finish my degree.....
Intents	My best friend had a really bitter break up. She constantly breaks down and gets lost in thought. I advised her to go to the gym. Now, she has been sweating at the gym daily. She is trying so hard so as to divert attention in tough times....
Emotions	... My mother sends a birthday card to my girlfriend, but not me. This birthday, I asked my girlfriend if I could rip the card as she had a bad day. She said yes, because she felt irritated when she received it.....
Emotions	.... there are these guys staring at me. I heard the cars stop behind me, I looked back so as to check if they were following me and I saw those guys coming towards me. I got inside the car and felt too anxious.

FIGURE 3.4: Dataset details: Samples extractions from Personal Narratives Corpus. The agent (ARG-0) and the purpose clauses (ARGM-PRP) are highlighted in red.

the annotators feel there is no clear intent or emotion for any instance). We find that the annotators agree with our extracted intents (Fleiss' $\kappa = 0.87$) and emotion (Fleiss' $\kappa = 0.90$) texts in 89% and 93% of the cases respectively.

3.1.2 Social Commonsense Knowledge

Though explicit motivation and emotion expressions are extracted by methods explained in the previous section, the implicit motives and emotions of characters expressed in sentences are not captured. To obtain those implicit

Personal Narratives Corpus	
#total narratives	439,408
#avg characters per story	2.02
#narratives w/ mappings	85,587
#sents w/ motives	167,256
#sents w/ emotions	318,872
% first-person motives	48.01%
% first-person emotions	58.26%
SBSCK Dataset	
#events w/ motives	103,357
#events w/ emotions	69,584
#unique social roles	586

TABLE 3.2: Statistics of Personal Narratives Corpus (top) and Search-based Social Commonsense Knowledge (SBSCK) dataset (bottom).

states, we: (a) exploit social commonsense knowledge (SCK) obtained from ATOMIC [Sap+19a] and ConceptNet [LS04] and (b) mine the web to augment more knowledge about the events from the personal narrative corpus. While ATOMIC contains inferential knowledge based on 24k short events, the knowledge from ConceptNet may not align with our requirements. For our purpose, we choose ConceptNet's relevant relations: /r/MotivatedByGoal, /r/CausesDesire, /r/Entails, /r/ Causes, /r/HasSubevent.

In our work, we posit that social roles (e.g., student, mother, boyfriend, etc.) provide extra information about the motives and emotions behind an action. The base events in knowledge sources – specifically ATOMIC contain typed markers (e.g., *PersonX*) where such information is lost. Therefore, we adopt web-based knowledge mining techniques to account for this extra information. The quality of such assertions may not be as high as well-curated knowledge collections like ATOMIC. However, they can act as an excellent source for pretraining our models. We refer to this as Search-based Social Commonsense Knowledge (SBSCK) data. Figure 3.5 shows samples from

this dataset that exemplifies how the same action could have different social role-related motivations. The following steps are involved in aggregating this dataset containing more social commonsense knowledge along with social role information: (a) process texts from our personal narratives corpus, (b) extract propositions from text using OpenIE tools, (c) perform a web search for plausible intents and emotions by attaching purpose clauses and feelings lexical units (explained earlier) and (d) finally, remove the poorly extracted facts using a simple classifier trained on some seed commonsense knowledge.

Search-based Social Commonsense Knowledge		
Social Roles	**Event Phrases**	**Motives**
Politicians	use social media	to woo voters
Activists		to create a movement
Police		to connect with residents and solve crime
Workers	gather around table	to solve business problems
Priests		to pray to god, share wine and bread
Friends		to share a meal, conversation

FIGURE 3.5: Dataset details: Samples motives related to specific social roles from Search-based Social Commonsense Knowledge (SB-SCK) dataset.

These steps involved in the data collection pipeline are described in detail below. We feed sentences from personal narratives to OpenIE tools which yield subject-relation-object triples. Next, we form a web search query q after normalizing [8] the triples and concatenating them with purpose clauses (for motivations) or feeling lexical units (for emotions). The query q is issued to the search engine, using its public API and enabling the spelling correction feature. We train a simple logistic regression using manually annotated seed

[8]For example, "clean a bedroom floor" is changed as "clean bedroom floor" using weak normalization and "clean floor" under strong normalization settings.

sets of search results to verify if they are valid candidates for knowledge extraction. We use the following features: average word embedding, number of words matched, exact match or approximate match, presence/number of stop words in mental state text, type of clause (purpose/feelings), and presence of entities. We use an N-V-OW representation scheme for words similar to [FDR19], where each word is categorized into: HeadNoun, FirstVerb, and OtherWords. Finally, we discard all results below a threshold score $\rho_{sck} < 0.35$. The data statistics are presented in Table 3.2. In Chapter 4, we primarily use the SB-SCK dataset to compute rich social event representation where the knowledge accumulated acts as a source of extracting pragmatics properties from the event text. However, there are severe shortcomings in applying them directly on stories due to the absence of context information. Using the computed sentence-level pragmatics-aware social event embeddings, we utilize the Personal Narratives Corpus for modeling motives and emotions of characters in the narrative given the story context. We describe this model in great detail in Chapter 5.

3.2 Identifying Climax & Resolution in Narratives

Figure 3.6 presents our data collection pipeline. The first stage in this pipeline is dedicated to ingesting posts from Reddit. To collect natural first-person stories, we rely on Reddit communities comprising user-generated textual accounts of happy events, long-standing baggage, recent trauma, life experiences, adventurous encounters or guilt, and redemption episodes. To this end, we aggregate posts from two communities: /r/offmychest and /r/confession using the PushShift API [9]. The Pushshift API provides access to a database of all Reddit posts made since Reddit's launch as a social platform. We obtain $\sim 440,000$ posts from this step.

[9] https://pushshift.io/

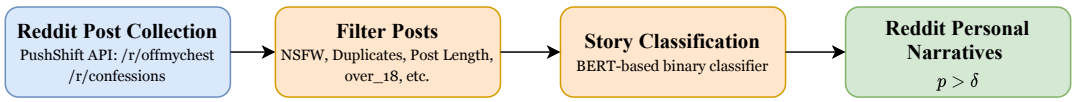


FIGURE 3.6: Illustration of our data collection pipeline.

Next, we filter the collected data to retain only those posts that do not contain tags like "[Deleted]", "NSFW" [10] or "over_18". Relying on Prince's definition [Pri12] of a minimal story to comprise a starting state, a state-changing event, and an ending state, we eliminate posts containing less than three sentences. The subsequent stages in the pipeline are detailed in the sections below.

3.2.1 Story Classifier

The aggregated data consists of a wide variety of contents, some of which do not qualify as personal narratives. In order to separate such non-narrative content from the collected data, we develop a story classifier that takes textual content as input and predicts the likelihood of the input text being a story.

Story vs. Non-Story Dataset We gather a diverse collection of first-person blog text drawn randomly from the Spinn3r Blog Dataset containing everyday situations [GS09]. Consistent with our filtering approach for Reddit posts, we follow a similar length criterion and sample $\sim 1,500$ blog posts. Further, we randomly selected $\sim 1,500$ texts from our Reddit posts corpus. Together, we obtain a total of $\sim 3,000$ posts to be annotated by MTurk workers.

For each post, annotators were instructed to read the textual content and choose one among the three labels: Story, Non-Story, or Unsure [Gor+13]. We define these categories as follows – (a) Story: Non-fictional narratives that

[10] NSFW – not safe for work

Models	P	R	F1
Sem. Triplet [Cer+12]	0.64	0.47	0.46
VerbNet [EF17]	0.71	0.66	0.68
HAN [Yan+16]	0.75	0.73	0.74
BERT [Dev+18]	0.81	0.78	0.79

TABLE 3.3: Performance of our BERT-based story classifier on the annotated dataset.

people share with each other about their own life experiences. They contain a sequence of causally or temporally related events with the narrator being a participant; (b) Non-Story: Texts that are not primarily personal stories or don't give an account of past events. They may or may not contain texts from the first-person point of view but include opinion pieces, excerpts from news articles, recipes, technical explanations, facts, questions or some random discussion, personal advice, to list a few. When the annotators are uncertain about the right label, they are allowed to select the "Unsure" option. The three workers reached unanimous agreement on 76% of the cases. We use the majority vote when such an agreement is not reached. Of the 3,000 posts, 1,197 posts were tagged as "Story", 1,173 as "Non-Story" and remaining as "Unsure".

Model We introduce a story classifier that separates non-fictional narratives from the non-narrative textual content. Prior work has used feature engineering to extract features like Tf-Idf, Semantic Triplets, VerbNet & coreference resolution chain based character features [Cer+12; EF17] for this task. A work by Piper [Pip18] specifically used the linguistic aspect of the text to measure fictionality, i.e., distinguish works of fiction from non-fiction. In our work, we use a pretrained BERT model for our classification task. Given an input text, the goal is to predict if the text qualifies as a story or not. We formulate the input text as $T = \{S_1, S_2, ..., S_n\}$, where S_i is the i^{th} sentence of the text. Following [Dev+18], we tokenize the input text and concatenate

all tokens as a new sequence, $\{[CLS], S_1, [SEP], S_2, [SEP], ..., S_{n-1}, [SEP], S_n, [SEP]\}$, where $[CLS]$ is a special token used for classification and $[SEP]$ is a delimiter. Each token is initialized with a vector by summing the corresponding token and position embedding from pretrained BERT, and then encoded into a hidden state. Finally, we get $[H_{[CLS]}, H_{S_1}, H_{S_2}, .., H_{S_n}, H_{[SEP]}]$ as an encoding output. We concatenate the $[CLS]$-token representation from the last four layers of the model for our classification task. We apply two linear layers on top of the concatenated output representation with a sigmoid activation function at the final linear layer. We optimize the binary-cross entropy loss and choose the model with the least loss on the validation set as our final story classifier. We evaluate this model on the held-out test set.

In Table 3.3, we report the F_1 score, and compare our approach to other baselines. The best performing model achieves an F_1-score of 0.79. Finally, we feed the Reddit posts to the trained story classifier and obtain a probability score, p, that indicates the likelihood of the post is a story. Furthermore, to increase the reliability of our data, we discard all those posts with a probability score lesser than a chosen threshold δ, i.e., $p < \delta$. In our work, we set δ to 0.75. This procedure yields a total of 63,258 stories, referred to as *Reddit Personal Narratives* dataset.

3.2.2 Annotation

Here, we explain the annotation process involved in constructing our manually annotated STORIES dataset. This dataset contains a total of 2,382 Reddit personal narratives, comprising 42,614 sentences. Table 3.4 shows the descriptive statistics of our dataset.

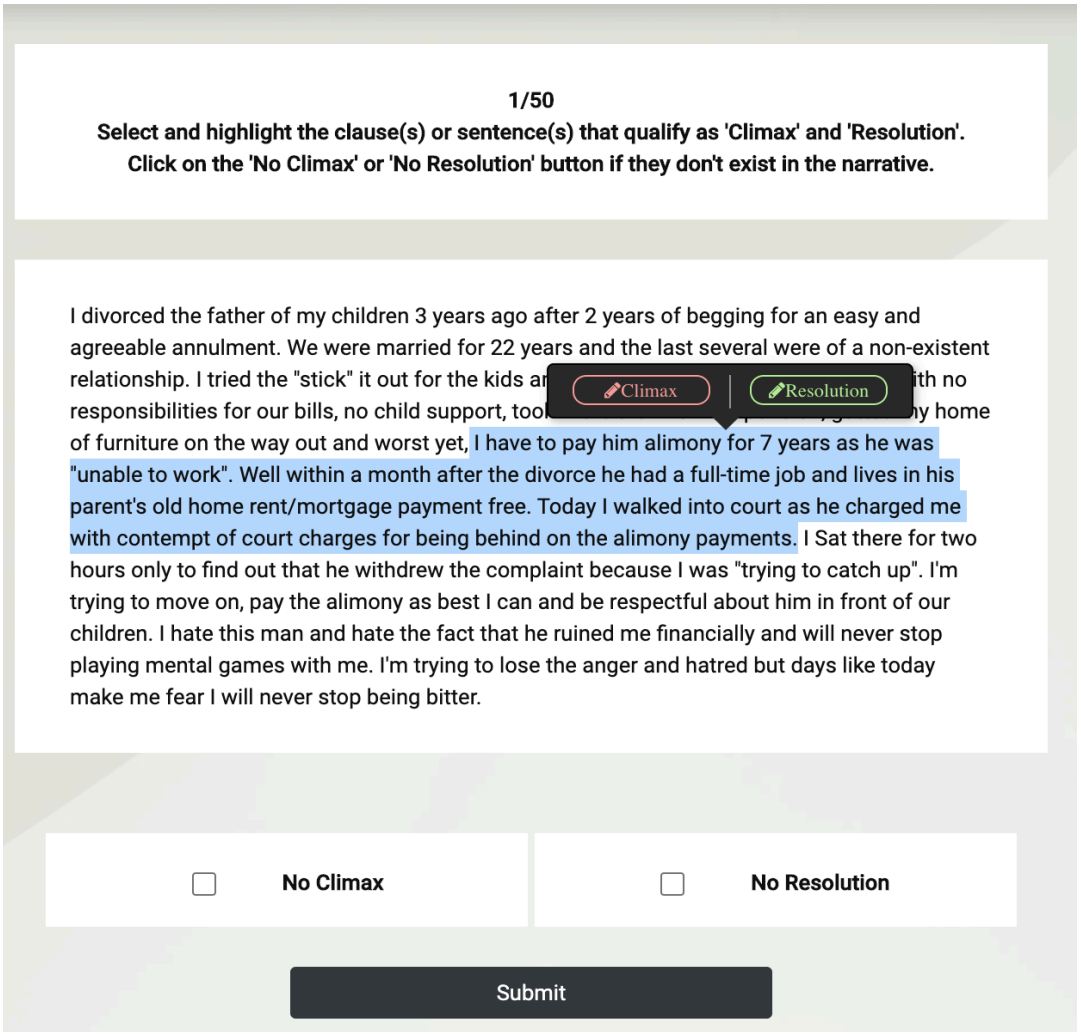


FIGURE 3.7: Sample Page from our user interface for annotation containing options to (a) highlight text and tag them as climax/resolution or (b) choose checkboxes – "No Climax" or "No Resolution", if annotators feel there is no climax and resolution.

Setup

We created a user interface for MTurk workers to make the annotation procedure convenient for capturing key elements of the narrative structure – climax and resolution. Towards formalizing and describing our annotation

scheme, proper guidelines were provided for the annotators. These guidelines include: (a) Definitions for both climax and resolution, (b) General annotation directions involving color schemes for highlighting the narrative elements, and (c) Select examples of personal narratives with colored highlights of identified narrative elements. The user interface allows the workers to highlight parts of the text that qualify as climax and resolution using red and green colors, respectively. Each worker has presented a sampled text from Reddit personal narrative corpus. Additionally, the workers are provided with an option of selecting checkboxes: "No Climax" or "No Resolution". This caters to those personal stories that don't contain a climax or resolution. Figure 3.7 provides an example of a page from our user interface for annotation purposes. A personal story is shown to the annotator, and options to highlight and tag them as climax and resolution are provided.

Dataset Statistics	
#Total Narratives	63,258
#Annotated Narratives	2,382
#Total Sentences	42,614
#Climax Sentences	5,173
#Resolution Sentences	4,502

TABLE 3.4: Statistics of our annotated STORIES dataset.

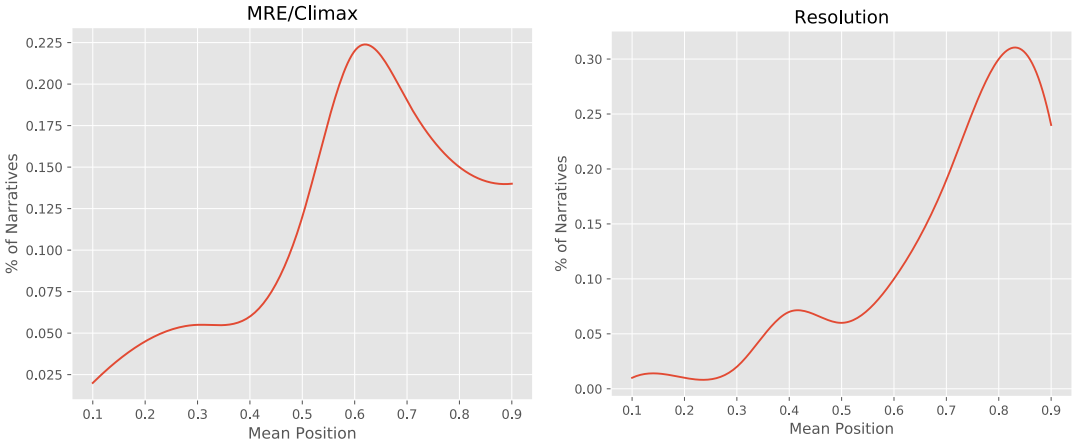


FIGURE 3.8: Distributions of mean climax & resolution sentence positions.

Agreements

Once the data is collected using our annotation setup, we measure the inter-annotator agreement (IAA) at both the sentence and span-level. For sentence-level agreement, we use the following metrics: (i) Cohen's kappa (κ) [Coh60], average pair-wise kappas are computed, (ii) mean annotation distance ($\mathcal{D}$), i.e., the distance between two annotations for each category, normalized by story length [PKL19]. Following [SRW08], we compute two measures for the text span agreement: (i) exact agreement in which the text spans are expected to match fully; and (ii) relaxed (lenient) matching in which the overlap between spans is considered as a match; and agreement naturally increases as we relax the matching constraints.

Sample Annotations
I've always wanted to have my own paycheck since I started high school. I've felt pathetic just asking my parents for money for the movies or for a school event or something like that. For the last few months of 2019 through March of this year, I've been applying to part-time jobs and getting nothing. It is so tiring. Getting rejected every time has brought me down to a significant low. But I finally got a job at the restaurant my dad works at as a bus boy and I get my first paycheck tomorrow. I'm so proud of myself. Something I can call my own.
It was snowing heavily and I had an instinct that something was amiss. I got a text message that my friend met with an accident. I was shocked. Thankfully, She's fine and so is everyone else who was in the car. But holy shit just seeing her message saying that she'd been in a car accident scared tf out of me. The instinct I had turned true. In a way, I'm worried about it and concerned how this will turn out. I'm still not quite over it, like idk why I still feel so weird and upset, but I realize shes okay and fine. I'm expecting to talk to her today evening. I hope talking to her might make me feel a lot better.
My co-worker has worked with us for a year now. We all just worked with her over the weekend. She had a dark sense of humor. She always joked about how life wasn't worth living during her shifts. And right on Monday she killed herself and was just gone. I'm totally broken. None of us are aware of how to respond to it. I really don't know what to make of it or how to process it. She was way too young to leave us. She never thought about us in her final moments. I'm still not out of the shock and struggling to get over this.

FIGURE 3.9: Sample annotations of climax (Red) and Resolution (Green) by one of the annotators.

Metric	Climax	Resolution
Span-level Analysis		
Exact Agreement (F1)	0.524	0.665
Relaxed Agreement (F1)	0.687	0.772
Sentence-level Analysis		
Percentage Agreement	0.736	0.807
Cohen's Kappa (κ)	0.651	0.769
Mean Annotation Distance ($\%\mathcal{D}$)	1.764	1.590

TABLE 3.5: Sentence and Span level inter-annotator agreement

Analysis

We study the appearance of climax and resolution sentences by estimating their mean position normalized by the story length. We present the distribution of the position of both the structural elements in Figure 3.8. While the expected average position for the climax (0.61) coincides with the peak, we observe that the resolution contents occur later in the story. Table 3.5 shows the sentence and span-level IAA measures for each narrative element. We observe that substantial agreement is achieved for both the climax and resolution. Clearly, the sentence-level analysis produces higher reliability scores than span-level measures. We attribute this to granularity, where the annotators marked expressions within the sentences in a close neighborhood with slightly different boundaries. Moreover, we obtain higher agreement values for resolution than the climax.

We analyze the discrepancies in the annotated data to gain insights into the potential challenges in the annotation process. For the climax, we note that the annotators get confused with sentences that involve events contributing to rising action (or Labov's complicating action), eventually culminating in a climax (or Labov's MRE). Further, we observe that the differences are more nuanced in many instances and hence harder to detect reliably. Though we achieve higher agreement on the resolution category, the annotation gets less accurate with ambiguities in resolution and aftermath/endings,

especially when narratives don't have a clear resolution. Interestingly, the annotators are able to discern between the two interest categories despite the high cognitive load and complexity involved in detecting them from unstructured user-generated content. Figure 3.9 displays sample annotations (e.g. multi-sentence or non-contiguous highlights; no resolution) from our STORIES dataset.

Chapter 4

Learning Knowledge-Enriched Social Event Representation

4.1 Introduction

Everyday life comprises the ways in which people typically act, think, and feel on a daily basis. Our life experiences unfold naturally into temporally extended daily events. The event descriptions can be packaged in various ways depending on several factors like speaker's perspective or the related domain. Interpretation of event descriptions will be incomplete without understanding multiple entities involved in the events and even more so when the focus is primarily on "social events", i.e., events explaining social situations and interactions. Therefore, a social event representation model must capture the semantic properties from the event text description and embed salient knowledge that encompasses the implicit pragmatic abilities. Early definitions of pragmatic aspects refer to the use of language in context; comprising the verbal, paralinguistic, and non-verbal elements of language [Ada+05]. Contemporary definitions have expanded beyond just communicative functions to include behavior that includes social, emotional, and communicative aspects of language [Ada+05; Par+17].

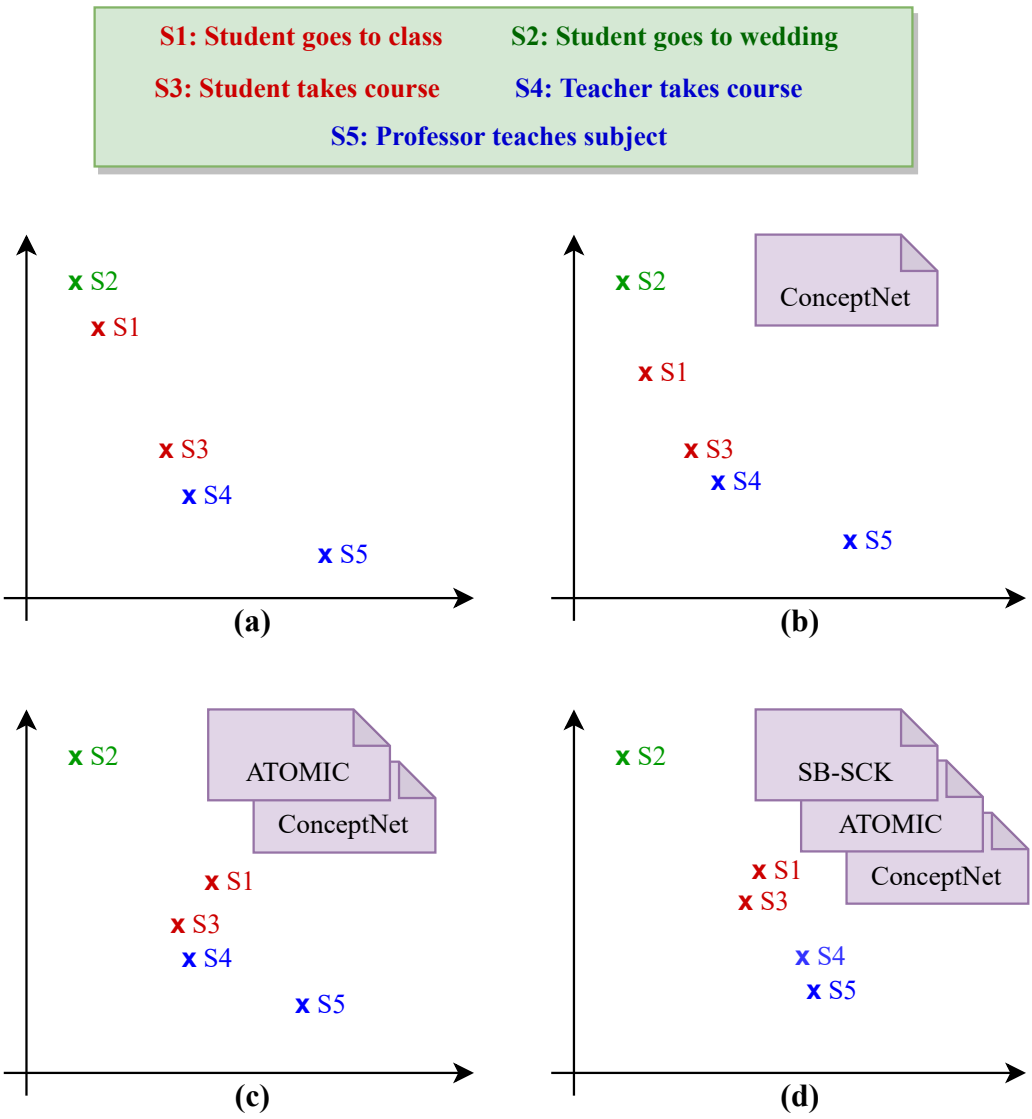


FIGURE 4.1: Illustration of functioning of our representation learning approach that produces rich social event embeddings. Event texts are given in the top green box. With more knowledge, social event embeddings move beyond high lexical overlap [shown in (a)] and learn to integrate semantic and pragmatic properties [shown in (b), (c)] of event texts along with social role information [shown in (d)].

Moving away from the extensively studied speech acts, we analyze characteristics that reflect how a person behaves in social situations and how social contextual aspects influence linguistic meaning. In the context of event representations, the pragmatic properties can specifically refer to the human's inferred implicit understanding of event actors' intents, beliefs, and feelings or reactions [Woo76; HN78].

Understanding the pragmatic implications of social events is non-trivial for machines as they are not explicitly found in the event texts. Prior studies [Din+14; Din+15; GWC16; WBC18] often extract the syntactic and semantic information from the event descriptions but ignore the pragmatic aspects of language. In this work, we address this shortcoming and aim to (a) disentangle semantic and pragmatic attributes from social event descriptions and (b) encapsulate these attributes into an embedding that can move beyond simple linguistic structures and dispel apparent ambiguities in the real sense of their context and meaning.

Towards this goal, we propose to train our models with social commonsense knowledge about events focusing specifically on intents and emotional reactions of people. Such commonsense understanding can be obtained from existing knowledge bases like ConceptNet [SCH17], Event2Mind/ATOMIC [Sap+19a; Ras+18a] or by collecting more noisy commonsense knowledge using data mining techniques. We, therefore, leverage these knowledge assertions aggregated from multiple sources to enable semantic and pragmatic enrichment of social event representations. One of the shortcomings of the dataset like ATOMIC is that the tokens referring to people are often replaced with a 'Person' marker. The social role information (e.g., student, mother, teacher, etc.) can significantly change the meaning of the event description and its interpretation as a whole. The motivation and emotional reaction associated with the same event can vary depending on the social role information. Figure 4.1 presents a sample functioning scenario producing incrementally richer social event embeddings. As the model gains more knowledge from different sources, it learns to discern events based on semantic and pragmatic properties, including social roles. For example, "Student takes course", and "Teacher takes course"have significant lexical and semantic relatedness. However, the social role information introduced by our own in-house aggregated SB-SCK dataset (as explained in Chapter 3) enhances the

representation learned from social event texts as depicted in Figure 4.1(d).

In this work, we develop a representation learning approach that learns to embed social event text at syntactic, semantic, and pragmatic levels. Our model utilizes a set of knowledge assertions from various domain sources to effectively integrate pragmatic attributes to interpret the real sense of events beyond lexical overlap or shallow semantic representation. Our contributions are as follows:

- We adopt a representation learning approach to disentangle the representation learned from event text into pragmatic and non-pragmatic properties and effectively consolidate the social commonsense knowledge from multiple domain sources and generate a semantically & pragmatically enriched social event embedding.
- We evaluate our models primarily on four different tasks: (a) intent-emotion prediction for event texts based on the social commonsense knowledge aggregated from different domains, (b) event similarity task using hard similarity dataset [Din+19; WBC18], (c) paraphrase detection using Twitter URL corpus [Lan+17a], and (d) social commonsense reasoning task using SocialIQA [Sap+19b] dataset.

4.2 Problem Formalization

Formally, we assume that our learning framework has access to streams of social commonsense knowledge data obtained from n different domains, denoted by $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n\}$. We denote j^{th}-input free-form event text in i^{th}-domain as $x_j^{(i)} = [w_1, w_2, ..., w_L]$. Here, $w_{(.)}$ refers to the tokens in the event text. Data from each domain source contains source-specific textual descriptions of social situations and their intuitive commonsense information such as intents and emotions. Training samples, drawn from a domain

dataset $\mathcal{D}_i$, could contain either a significant overlap or a completely new set of knowledge when compared with the previously processed domains $\mathcal{D}_{1:i}$. Given such a setup, we aim to generate richer social event representations using our representation learning framework.

4.3 Datasets

For our representation learning task, we aggregate social commonsense knowledge data[1] from various domain sources. This knowledge contains details about pragmatic aspects like intents and emotional reactions.

4.3.1 Social Events Dataset

Different domain sources of social commonsense knowledge used for training our social event representation model are explained as follows.

ATOMIC dataset consists of inferential knowledge based on 24k short events covering a diverse range of everyday events and motivations. Though each event contains nine dimensions per event, the scope of this work will be limited to intent and emotions as our inferential pragmatic dimensions.

CONCEPTNET knowledge base contains several commonsense assertions. For our purpose, we choose ConceptNet's relevant relations: /r/MotivatedByGoal, /r/CausesDesire, /r/Entails, /r/Causes, /r/HasSubevent. We convert triples in the dataset into template form.

SB-SCK As explained in Chapter 3, search-based social commonsense knowledge dataset contains additional social role information(e.g., student, mother,

[1]The project details about future data/code releases or any updates will be available at https://pralav.github.io/lifelong_eventrep?c=10

Search-based Social Commonsense Knowledge		
Social Roles	**Event Phrases**	**Motives**
Politicians	use social media	to woo voters
Activists		to create a movement
Police		to connect with residents and solve crime
Workers	gather around table	to solve business problems
Priests		to pray to god, share wine and bread
Friends		to share a meal, conversation

SB-SCK Dataset	
#events w/ motives	103,357
#events w/ emotions	69,584
#unique social roles	586

FIGURE 4.2: **Left**: Samples from Search-based Social Commonsense Knowledge (SB-SCK) dataset with highlighted motivations for social roles, **Right**: Statistics of SB-SCK dataset.

teacher, worker, etc.) that provide details about the social context and its inferred motives and emotions behind actions specified in the events (as shown in Figure 4.1, 4.2, we adopt web-based knowledge mining techniques for capturing these knowledge assertions. Figure 4.2(Left) shows samples from this dataset indicating how the same action could have different social role-related motivations. We refer to this as Search-based Social Commonsense Knowledge (SB-SCK) data. Figure 4.2(Right) presents the data statistics.

For data from each of the above domain sources, we sample free-form event text, paraphrase, intent, emotional reactions, and negative samples of paraphrases, intents, and emotional reactions. Based on the annotated labels for motivation (Maslow's) and emotional reactions (Plutchik) in STORYCOMMONSENSE data, we run a simple K-Means clustering on the open text intent data. We identify five disjoint clusters on each of the three domains and map them to those categories. We use these categories so that different types of data are sufficiently represented in our train, valid, and test sets in this work.

4.3.2 Paraphrase Datasets

We use random samples of parallel texts from paraphrase datasets like PARANMT-50M corpus[WG17] and Quora Question Pair dataset [2]. These paraphrase datasets are primarily used for pretraining our model. We also produce paraphrases of free-form event texts in our dataset using a back-translation approach [Iyy+18]. We used pretrained English↔German translation models for this purpose.

4.4 Framework

Our goal is to learn distributed representations of social events by incorporating pragmatic aspects beyond shallow event semantics. Moving away from conventional supervised multi-task classification-based learning approaches, we focus on a representation learning approach that enables us to adapt and learn a social event embedding model. The motivation for learning such representations is to uncover latent information at syntactic, semantic, and pragmatic levels by exposing the model to the knowledge about implicit mental states of event actors'. This knowledge is obtained from various domain sources and can effectively guide the modeling of complex social events and extract the meaning of the events beyond shallow features. In this section, we will explain various components of our modeling framework.

4.4.1 Social Event Representation

Given an input event text description, the core idea is to encode the free-form event text and decompose the ensuing representation into pragmatic (implied emotions and intents) and non-pragmatic (syntactic and semantic

[2]https://www.kaggle.com/c/quora-question-pairs/data

information) components. Eventually, we combine these decomposed representations to obtain an overall event representation and apply it in different downstream tasks.

Encoder

The input to our model is a free-form event text description from i^{th} domain, $x_j^{(i)} \in \mathcal{D}_i$. This free-form event text contains a sequence of tokens, $x_j^{(i)} = [w_1, w_2, ..., w_L]$, where each token $w_{(\cdot)}$ is obtained from an input vocabulary $\mathcal{V}$. The model encodes the input event text $x_j^{(i)} \in \mathcal{R}^{L \times d_X}$ in multiple steps. First, we construct a context-dependent token embedding using a context embedding function $\mathcal{G} : \mathcal{R}^{L \times d_X} \mapsto \mathcal{R}^{L \times d_H}$, where d_X and d_H refer to the embedding and hidden layer dimensions respectively. Following this encoding step, we incorporate pooling or projection function, $\mathcal{G}_{p\bar{p}}$: $\mathcal{R}^{L \times d_H} \mapsto \mathcal{R}^{3 \times d_H}$, that transform event text from context-dependent embedding space into pragmatic and semantic space. More specifically, we produce latent vectors for intents (h_I), reactions (h_R) and non-pragmatic (h_N) information. Finally, we combine the latent vectors h_N, h_I, h_R using a simple feed-forward layer, $\mathcal{G}_C$: $\mathcal{R}^{3 \times d_H} \mapsto \mathcal{R}^{d_H}$, to produce a rich social event embedding, h_C. Given positive and negative examples of intents, emotional reactions and paraphrases associated with the input event text, we learn to effectively sharpen each of these embeddings h_I, h_R and using metric learning methods.

For the sake of brevity, we drop the domain index i and the sample index j in this section. These encoding steps are summarized as:

$$H_e = [h_1, h_2, ..., h_L] = \mathcal{G}([w_1, w_2, ..., w_L]) \tag{4.1}$$

$$h_I, h_R, h_N = \mathcal{G}_{p\bar{p}}(H_e) \tag{4.2}$$

$$h_C = \mathcal{G}_C(h_I, h_R, h_N) \tag{4.3}$$

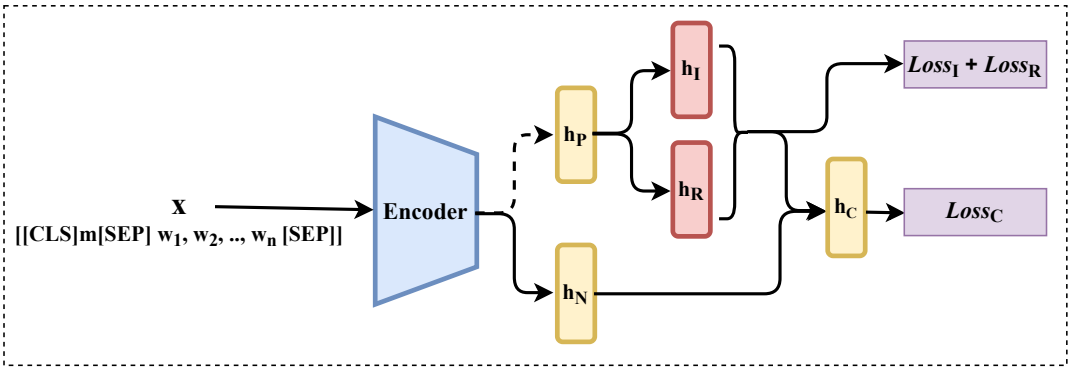


FIGURE 4.3: **Left:** Illustration of our Social Event Representation Model.

We denote this multi-step encoding process resulting in h_I, h_R, h_C as a function $\mathcal{G}_{event}$. Now, we experiment with the following text embedding techniques as our context embedding function ($\mathcal{G}$):

BiGRU : We use a recurrent neural network to encode the input sequence. More precisely, we choose a gated recurrent network (GRU) over LSTM as they achieve comparable performance levels with lesser computational resource requirements. Using bidirectional GRUs, we obtain forward ($\overrightarrow{h_t}$) and backward hidden states ($\overleftarrow{h_t}$) of the input sequence. We concatenate these forward and backward hidden states at each timestep $t \in \{1, 2, ..., L\}$ to get an overall latent vector representation (H_e) of the input event text. This is computed as:

$$H_e = [\overleftrightarrow{h_1}, \overleftrightarrow{h_2}, ..., \overleftrightarrow{h_L}] = \overleftrightarrow{\text{GRU}}(x) \tag{4.4}$$

$$\overrightarrow{h_t}, \overleftarrow{h_t} = \overrightarrow{\text{GRU}}(w_t, \overrightarrow{h_{t-1}}), \overleftarrow{\text{GRU}}(w_t, \overleftarrow{h_{t-1}}) \tag{4.5}$$

$$\overleftrightarrow{h_t} = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{4.6}$$

BERT We employ BERT [Dev+18], a multi-layer bidirectional Transformer-based encoder, as our context embedding method $\mathcal{G}$. We fine-tune a BERT model that takes attribute-augmented event text $x = [CLS]\ m\ [SEP]\ w_1, ..., w_L\ [SEP]$ as input and outputs a powerful context-dependent event representation H_e.

The attribute $m \in \{xIntent, xReact, xNprag\}$ refers to special tokens for intents, reactions and non-pragmatic aspects. Special tokens $[CLS]$ and $[SEP]$ are commonly added as first and last tokens. Formally, we define it as:

$$H_e = [h_1, h_2, ..., h_L] = \text{BERT}(x) \tag{4.7}$$

In our default case, our $\mathcal{G}_{p\bar{p}}$ function is the output embedding of $[CLS]$ token associated with their respective attribute-augmented input. In cases where input event text is not augmented with attribute special tokens, we apply pooling strategies such as attentive pooling (AP) and mean (MEAN) of all context vectors obtained from the previous encoding step $\mathcal{G}$. We obtain h_I, h_R, h_N based on these techniques. Depending on the type of context embedding function, we refer our multi-step event text encoder, $\mathcal{G}_{event}$, as EVENTGRU or EVENTBERT.

Objective Loss

Using positive $\{u_I^p, u_R^p, u_C^p\}$ and $N-1$ negative $\{u_I^n, u_R^n, u_C^n\}$ examples of intents, emotions and paraphrases associated with the event texts, we calculate N-pair loss, $\mathcal{L}_v(h, z^p, \{z_k^n\}_{k=1}^{N-1})$, to maximize the similarity between the representation of positive examples (z_v^p) and the computed embeddings (h_v). Here, z_v^e is computed using a transformation function f_v as: $z_v^e = f_v(u_v^e)$, where $v \in \{I, R, C\}$ and $e \in \{p, n\}$. Thus, our loss function is devised as:

$$\mathcal{L}_\mathcal{T} = \frac{\beta_D}{2} \cdot (\mathcal{L}_I + \mathcal{L}_R) + \beta_E \cdot \mathcal{L}_C \tag{4.8}$$

where $\mathcal{L}_I, \mathcal{L}_R$ are used to learn disentangled pragmatic embeddings (intent and emotion), $\mathcal{L}_C$ is intended to jointly embed semantic and pragmatic aspects to produce an overall social event representation. β_D, β_E are loss coefficients that weigh the importance of disentanglement loss and an overall joint

embedding loss. These coefficients are non-negative and they sum to 1.

4.5 Training

Since our model involves metric learning, hard negative data mining is an essential step for faster convergence and improved discriminative capabilities. However, selecting too hard examples too often makes the training unstable. Therefore, we choose a hybrid negative mining technique where we choose a few semi-hard negatives examples [HBL17] and combine them with random negative samples to train our model effectively. Usually, it is also unclear what defines "good" hard negatives [HBL17].

In our work, we define a heuristic objective by weighing samples based on two factors: (i) word overlap or similarity in embedding space of the event text and (ii) intent and emotion free-form text or categories based on STORYCOMMONSENSE data. More specifically, given an event text as an anchor and a positive intent text based on a ground truth motivation category, we mine negative instances for intent as follows: (a) choose random text samples associated with a motivation category that is different from that of the positive example but closer in the embedding space or word overlap, (b) choose random text samples within the same motivation category but with different emotion category. We repeat this process for drawing negative instances related to emotions. For paraphrases, we consider few examples with significant word overlap while the rest are randomly chosen samples. Since N-pair loss function allows for faster convergence and alleviates challenges in hard mining strategy, we utilize N-pair loss as our objective function. N-pair loss helps lessen the sensitivity of triplet loss function to the choice of hard triplets. This is done by pushing away multiple negative examples jointly at each update. Before using these negative intent/emotion samples, we pretrain our model with paraphrase data to capture different forms of conveying

Methods	Intents	Emotions
ConvKB	58.87	71.05
NTN	59.19	71.58
ERNIE	64.37	74.46
EventGRU	60.06	70.61
EventBERT	**70.03**	**79.96**

TABLE 4.1: Evaluation results on the held-out test set. We report the accuracy (%) scores for different baselines. Boldface indicates the best accuracy scores for a particular category (intents/emotions).

the same semantic content. For this pretraining, we sample examples from our paraphrase dataset explained in 4.3.2. Finally, we pre-train our model with paraphrase data and fine-tune it using the examples obtained from hard negative mining for intents, emotions, and paraphrases. For our training, the learning rate is set to 0.0001, the number of training epoch is 20. We conduct a study by assigning different values for loss coefficients, β_D, β_E, and explain their results in Section 4.6.1.

4.6 Experiments

In this section, we experiment with our learned social event representations on different NLP tasks: intent-emotion prediction, paraphrase detection, Social IQA reasoning, and event similarity. While we utilize the intent-emotion prediction task for evaluating our continual learning setup, we establish the richness of our social event embeddings using the remaining downstream tasks.

4.6.1 Intent-Emotion Prediction

We evaluate our trained models on a held-out test set across the different domains in our aggregated dataset (see Section 4.3). By default, we use EVENTBERT as our multi-step encoder.

Setup

In this work, we compare against different baseline neural network approaches applied in the past for knowledge embedding on our training set. We report the test set performance on predicting intents and emotions. The baselines are listed below.

- **ConvKB** [Ngu+18] We train a variant of this CNN-based method by feeding event text as triples to this model. We evaluate this method by applying linear layers on top of the output feature vector.
- **NTN** [Soc+13; Din+15; Din+19]: This method utilizes neural tensor network to perform semantic composition of event arguments. The bilinear tensors are explicitly applied to model the relationship between event actor and their actions. We apply linear layers on top of the final representation.
- **ERNIE**[3] [Zha+19]: We utilize a variant of this model consisting of a Transformer-based textual encoder and a knowledge encoder that fuses knowledge and textual information into a united feature space. We use this model and apply a linear layer on top of the representation for intent and emotion prediction.
- **EventGRU**: This is our model variant with a GRU-based encoding strategy.
- **EventBERT**: This is our complete model in default settings. We compare this proposed approach over the above baselines.

Empirical Results

Table 4.1 reports the accuracy scores for the intent and emotion prediction task. We note that our EventBERT model outperforms all the other baselines

[3]https://github.com/thunlp/ERNIE

Pooling Strategy	Intents	Emotions
AP	65.50	76.13
MP	67.28	76.69
CLS	**68.56**	**78.48**

TABLE 4.2: Results of our ablation study related to the pooling strategy on the held-out validation set.

in this task. The closest performing model is ERNIE which is a transformer-based model that integrates knowledge representation with textual information. Despite ERNIE's improved performance on many knowledge graph-related tasks, our EventBERT records the best performance on this task. We intuit that the main reason for this performance lies in the advantage of jointly training on the aggregated datasets with the ability to disentangle for pragmatic properties effectively. Moreover, we note that models attain the best performance by permuting the training set across various domain sources instead of sequentially training separately on individual datasets.

Ablation Study

We conduct an ablation study by analyzing various model configurations related to: (a) pooling: attribute-augmented input (CLS), Mean Pooling (MP), and Attentive Pooling (AP), and (b) loss co-efficient: β_E. For each pooling strategy, we did compare the model performance for different values of β_E. However, we report only the best performing configuration for each pooling strategy towards predicting intents and emotions. Table 4.2 shows the results of different pooling strategies for intent and emotion prediction task.

Additionally, we measure the effect of β_E in the prediction of intents. As shown in Figure 4.4, the model performs significantly better for lower values β_E as more weight is assigned for the disentanglement of pragmatic aspects. Since we are evaluating here precisely to predict intents, the disentangling

coefficient plays a critical role. We do have models trained using these different coefficients and use the representations for downstream tasks. Though the best performing model may vary depending on the task under consideration, we observe that a balanced loss function with $\beta_E = 0.5$ allows for consistently good performance in both intent-emotion prediction (Figure 4.4) and hard similarity tasks (see Section 4.6.2). Despite other hyperparameters, changes to β_E determine the importance of incorporating semantic or pragmatic information in the ensuing event embedding.

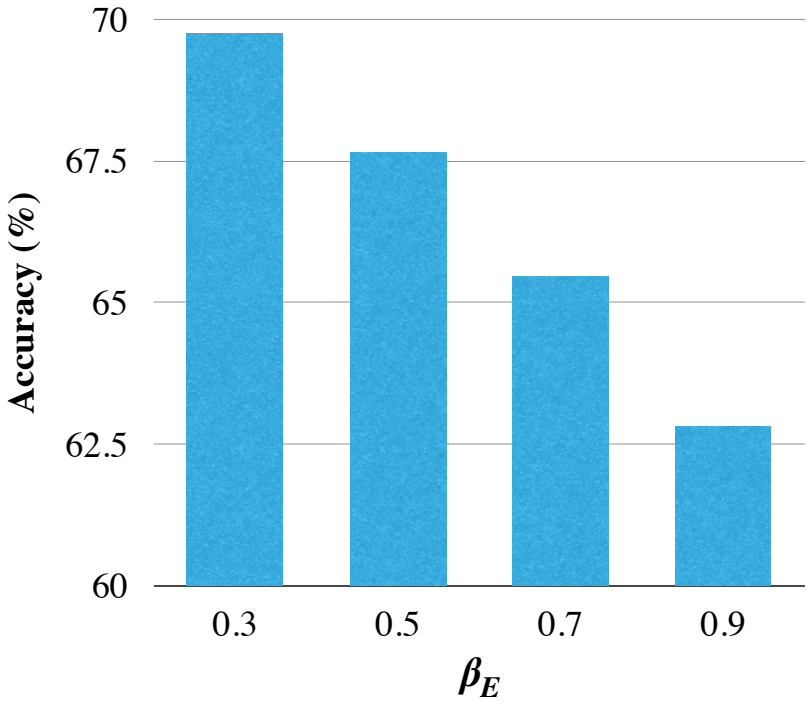


FIGURE 4.4: Results of our ablation study on a held-out validation set. ***Acc*** scores (%) to measure the effect of β_E in predicting intents.

4.6.2 Hard Similarity Task

By following the work of Ding et al. [Din+19], we evaluate our social event representation on an extended dataset of event pairs containing: (a) similar event pair having minimum lexical overlap (e.g., people admired president/ citizens loved leader) (b) dissimilar event pair with high lexical overlap (e.g.,

Models	% Acc.
KGEB	50.09
NTN + Int	58.83
NTN + Int + Senti	64.31
EVENTBERT$_{\beta_E=0.3}$	66.19
EVENTBERT$_{\beta_E=0.5}$	**71.23**
EVENTBERT$_{\beta_E=0.7}$	69.79

TABLE 4.3: Evaluation results on the combined hard similarity dataset.

people admired president/ people admired nature). A good performance in this task will ensure that similar events are pulled closer to each other than different events. Combining hard similarity datasets from [Din+19] and [WBC18], the total size of this balanced dataset is 2,230 event pairs. Using our joint embedding h_C for an event text and triplet loss setup, we compute a similarity score between similar and dissimilar pairs. The baselines include: Knowledge-graph based embedding model (KGEB) [Din+16], Neural Tensor Network (NTN) and its variants augmented with ATOMIC dataset based embeddings (Int, Senti) [Din+19]. We report the model's accuracy in assigning a higher similarity score for similar pairs than dissimilar pairs. Table 4.3b shows that our model outperforms the state-of-the-art method for this task.

4.6.3 Paraphrase Detection

To assess the quality of our learned embedding, we present an evaluation on the paraphrase detection task. Given a sentence pair, the objective is to detect whether they are paraphrases or not. For each sentence pair (s_1, s_2), we pass them through our model and obtain their respective h_C, given by vectors (u, v). We concatenate these vectors (u, v) with the element-wise difference $|u - v|$ and fed to a feed-forward layer. We optimize binary cross-entropy loss. For evaluation purposes, we compare our model against baselines like BERT and ESIM [Che+16]. Trained on a subset of the dataset explained in

Models	% Acc
ESIM	84.01
BERT	87.63
EVENTBERT$_{0.5}$	88.23
EVENTBERT$_{0.7}$	**90.16**

TABLE 4.4: Accuracy scores (%) of different models on Twitter URL Paraphrasing corpus, TwitterPPDB. Subscript of EVENTBERT model indicates value of β_E.

Section 4.3.2, we choose an out-of-domain test dataset where samples stem from a dissimilar input distribution. To this end, Twitter URL paraphrasing corpus [Lan+17a], referred to as TwitterPPDB, is selected. This dataset contains sentence pairs from Twitter where tweets are considered paraphrases if they have shared URLs. We used a 3-month collection of paraphrases. Table 4.4 contains results of our evaluation. The results testify to the efficacy of our embeddings.

4.6.4 Social IQA Reasoning

We determine the quality of our latent social event representations by evaluating on a social commonsense reasoning benchmark – SocialIQA dataset [Sap+19b]. Given a context, a question, and three candidate answers, the goal is to select the right answer among the candidates. Since our social event embedding approach models particular pragmatic components like intents and emotions, we assume that our model will help score better on specific question types like 'motivations' and 'reactions'. Trained on a dataset of around 33k samples explained in [Sap+19b], BERT achieves state-of-the-art performance in this multiple-choice implementation setup. Following Sap et al.[Sap+19b], the context, question, and candidate answer are concatenated using separator tokens and passed to the BERT model. Additionally, we feed the context to our EVENTBERT model to obtain three embeddings h_I, h_R, h_C.

Models	Dev	Test
w/o Social Event Embeddings		
GPT2	63.3	63.0
BERT-base	63.3	63.1
BERT-large	**66.0**	**64.5**
w/ Social Event Embeddings		
BERT-base	65.1	64.0
BERT-large	**68.7**	**67.9**

TABLE 4.5: Accuracy scores (%) of different models on SocialIQA dev and test dataset. The best accuracy is indicated in boldface.

While the original work computed a score l using the hidden state of $[CLS]$ token, we introduce a minor modification to this step as:

$$l = W_5 \, tanh(W_1 \, h_{CLS} + W_2 \, r_{xIntent} + W_3 r_{xReact} + W_4 \, r_C) \tag{4.9}$$

where $W_{1:4} \in \mathcal{R}^{d_H \times d_H}$ and $W_5 \in \mathcal{R}^{1 \times d_H}$ are learnable parameters. Similar to [Sap+19b], triple with the highest normalized score is used as the model's prediction. We fine-tune BERT models using our new scoring function with social event embedding (denoted as "w/") and compare against baselines (like GPT/GPT2 [Rad+18]) without our event embeddings (denoted as "w/o"). We report the scores directly from the original work [Sap+19b]. Results in Table **??**a indicate that a simple enhancement procedure at the penultimate step can offer significant performance gains. Our findings suggest that our enhanced model performed well for question types like 'wants' and 'effects' that weren't explicitly modeled in our embedding model. This confirms that our pragmatics-enriched embeddings lead to improved reasoning capabilities.

4.7 Conclusion

Humans rely upon commonsense knowledge about social contexts to ascribe meaning to everyday events. This social commonsense knowledge may include a growing set of norms of behavior and pragmatic implications of the participants' actions in a given social situation. In this work, we introduce a representation learning approach for the effective representation of social events with the help of social commonsense knowledge assertions acquired from different domains. By incorporating social commonsense knowledge with our text encoding techniques, we learn rich embeddings of social events from their free-form textual descriptions. First, we sharpen the semantic and pragmatic aspects of social events using social commonsense knowledge and jointly capture the overall non-ambiguous meaning of the event text. Using an intent-emotion prediction task, we evaluate the learning setup based on a held-out corpus of social events obtained from multi-domain knowledge sources. By evaluating this held-out corpus of social events obtained from multiple domain sources, we establish that our model is able to outperform several baselines.

Experimental results on downstream tasks like event similarity, reasoning, and paraphrase detection tasks demonstrate our social event embeddings' capabilities. However, we note that the trained model might not encompass all the knowledge necessary to handle novel social situations involving cultural context as we don't model for that explicitly. More relevant knowledge assertions embodying cultural information can be helpful in such scenarios. Instead of training the model from scratch for growing knowledge, lifelong learning approaches for social event representation can guide the accommodation of new knowledge and promoting positive knowledge transfer to new domains [VR21a]. We hope that our work will motivate further exploration into lifelong representation learning of social events and advance

the research in inferring pragmatic dimensions from texts.

Chapter 5

Modeling Human Motives and Emotions from Personal Narratives

5.1 Introduction

Narratives are one of the most common yet powerful means of communication used to enhance engagement with people's issues and understanding of the social world. People share and consume them in a variety of ways to convey and make sense of their experiences. Theorists and researchers in a wide variety of fields like neuroscience, psychology, and narratology have long posited that narratives exert a powerful influence on social cognition by evoking mentalizing process [GW02; GPHL08; CSG98; CWC11]. Mentalizing is used to describe all kinds of reasoning about others' mental states, such as inferring other peoples thoughts, beliefs, attitudes, emotions, and motivations. Studies have argued that reading more stories in one's lifetime and analyzing characters' behavior in stories contributes to greater activation of mentalizing network [Mar18]. Therefore, comprehending narratives is key to understanding human agency.

In this work, we are specifically interested in uncovering certain aspects of the relationship between narratives and mentalizing [Fer+08; Mar11; Tam+16b; MG17]. We focus on developing computation approaches to model human

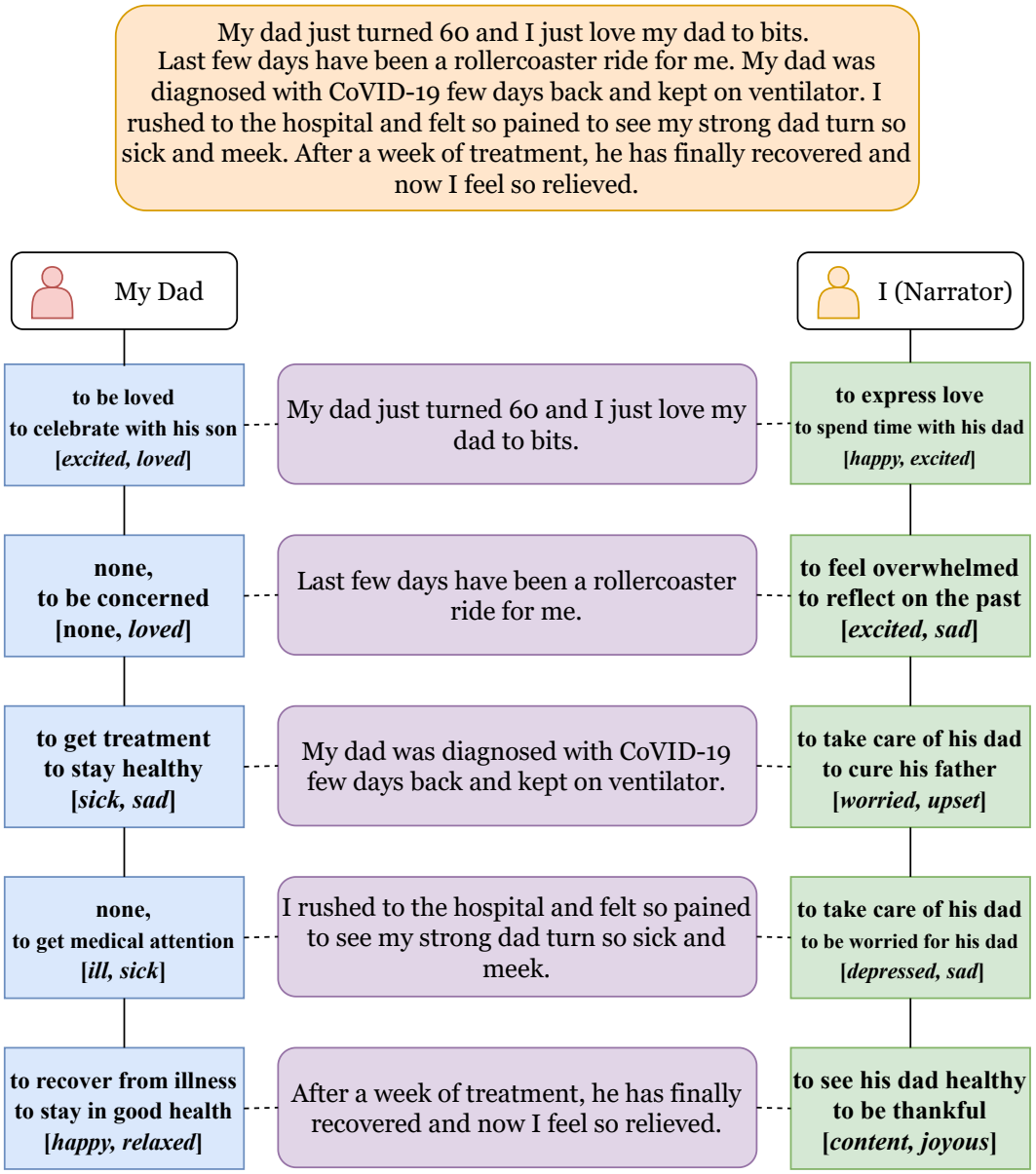


FIGURE 5.1: Sample personal narrative is shown on the top. It contains the motives and emotional reactions [*italics*] of different characters – dad and son (narrator) in the narrative.

motives and emotions from narratives containing explicit and implicit references to the characters' psychological states and their corresponding social contexts. To this end, different models of narrative analysis such as Labov's "evaluative devices" [GRDI10], or Lehnert's "plot units" [Leh81] have been proposed to track the mental states or affect states of the characters towards narrative understanding and summarization. A work by [Ras+18c; Ras+18a] focused on constructing a dataset comprising rich low-level annotations of

categories and textual explanations of motivations and emotional reactions of characters in five-sentence stories. By modeling character-specific contexts and pretraining on free-text responses, they provide benchmark results on this new resource. However, very limited work focuses on rich representation and generation of textual explanations of mental states, precisely motives, and emotional reactions. Also, there is tremendous scope for improvement in furthering the research towards imparting mentalizing capabilities for machines. Some of the key challenges in modeling human motives and emotions include: (a) lack of annotated data that captures explicit and implicit mental states of characters in narratives from different domains, (b) ability to track characters' mental state shifts continuously, and (c) effectively embed and generate their corresponding text explanations.

To tackle a subset of the aforementioned challenges, we resort to personal narratives from social media. Similar to a literary story, a personal narrative is likely to contain a beginning, middle, and end, where the middle typically presents a complication for the person, one that is resolved in some way by the ending. Similarly, it may convey information about goals, motives, thoughts, conflicts, emotions, and resolutions of people, including self or other people inside or outside their social circle [GW11; Abb20]. This makes them a practical resource for knowledge extraction and modeling. Since manual annotation is usually labor-intensive and expensive, we adopt a combination of web data mining and information extraction (IE) strategies to automatically extract and aggregate noisy expressions of motivations and emotions related to specific events in the text (applicable to different textual domains). This facilitates the acquisition of weakly-annotated data containing characters' motivations and emotions from personal narratives and social commonsense knowledge from the web. Figure 5.1 (top) shows a sample personal narrative from Reddit with character-specific explanations of intents and emotions behind every event in the narrative. Consider the sentence "I

rushed to the hospital...", the intent of the narrator ("I") is "to take care of his dad" or "to be worried for his dad". To produce such explanations, it is necessary to condition on the story context and social role because modifying them could significantly alter the meaning and their corresponding intent and emotional reactions behind the same action (e.g. "doctor rushed to the hospital" could have a different intent: "to attend to an emergency patient").

Thus, our goal is to (a) develop rich representations of mental states of humans grounded in intuitive theories of human psychology and commonsense knowledge, (b) generate textual explanations of mental states considering the prior context and social role information, and (c) harness transferability to downstream tasks. We, therefore, implement a Transformer-based encoder-decoder architecture, referred to as NEMO[1] to embed and explain characters' (or entities') mental states. To this end, we equip our model with components that: (a) enable pragmatic enrichment of narrative sentences using the aggregated knowledge and (b) track entities' mental states over time using an external memory module. Inspired by the ideas from cognitive science [GV10], these components can be perceived as analogous to certain characteristics of semantic and episodic memories. Thus, our contributions are as follows:

- Data collection of Personal Narratives [2] and Social Commonsense Knowledge containing weak-annotations of motivation and emotion text expressions.
- An end-to-end Transformer-based NEMO model augmented with modules that infuse social commonsense knowledge and dynamically track entities' mental states.

[1] Short for **N**arrative **E**ntity **M**ental m**O**del
[2] We will be making the data available soon.

- Trained on the aggregated weakly annotated data, we conduct experiments on the STORY- COMMONSENSE dataset [Ras+18c] under various evaluation settings. To exemplify our learned embeddings' transferability, we perform a simple evaluation on EMPATHETICDIALOGUES dataset.

5.2 Problem Setup

Formally, a story $\mathcal{S}$ consists of a sequence of T sentences $\mathcal{S} = [s^{(1)}, s^{(2)}, .., s^{(T)}]$ and a set of N entities/characters $\mathcal{E} = \{e_1, e_2, .., e_N\}$. We denote t^{th}-sentence containing L words as $s^{(t)} = [w_1^{(t)}, w_2^{(t)}, ..., w_L^{(t)}]$. Given an entity e_j, current story sentence $s^{(t)}$ and prior story context $s^{(<t)}$, we aim to generate mental state explanations of e_j, $\mathcal{Y}_m = [y_m^{(1)}, y_m^{(2)}, ..., y_m^{(T)}]$, related to mental state attribute $m \in \{xIntent, xReact\}$. Therefore, our approach models the conditional probability: $P(y_m^{(t)} | s^{(t)}, s^{(<t)}, y_m^{(<t)}, e_j, m)$.

5.3 Related Work

There has been a growing interest in developing computational models to model aspects of human behavior from day-to-day events or stories. Prior work by [GRI13] presented a system Aesop that builds on the idea of Lehnert's plot units [Leh81] and utilizes existing resources to predict affect states of characters in Aesop Fables. A line of work by [CGDI16; Rah+17] focused on modeling desire and fulfillment. This work considers five or fewer sentences to model the context of the desire expression and developed a logistic regression-based classifier for the desire fulfillment prediction task. There has been a recent body of research [Gui+17; Gho+17] that detects emotional stimuli in stories and generates text based on specific attributes like sentiment or affect states based on LIWC categories. One of the closest works in

this space is [Ras+18c]'s resource for character mental state tracking in short five-sentence commonsense stories. In our work, we develop automatic techniques to extract weakly-annotated mental state expressions with social role information being retained from personal narratives (a more natural setting) and propose a method to leverage social commonsense knowledge to generate and classify character motivation and emotion states efficiently. Further, we address the modeling challenges by incorporating social commonsense knowledge from social events and employing entity modeling for tracking the mental states of characters in the narrative.

Prior work in entity modeling is limited by their ability to track simple attributes, entity reference or specific physical properties of entities as in [Hen+16; Bos+17; Ji+17]. In this work, we focus on capturing the dynamics of the entities' previous motivation and emotional states. We achieve this by equipping our model using a memory module with operations involving decoder contextual hidden states. It is worth noting that models that incorporate entity-aware memory-based target-side context are a rarity. We intuit that employing attention mechanism over prior decoder states (target-side context) facilitates improved explanation generation by efficiently recording the motivation and emotion states.

5.4 NEMO: Our Proposed Model

Our overall objective is to learn character-specific embeddings of mental states – especially motives and emotional reactions, and produce their textual explanations by integrating external knowledge along with social role information, preceding narrative context, and mental state encodings. In this direction, we introduce a Transformer-based encoder-decoder architecture augmented with external memory modules that enable knowledge-enrichment

and dynamic state tracking of entities. Figure 5.2 provides an overview of our NEMO model. The prime components of our model include:

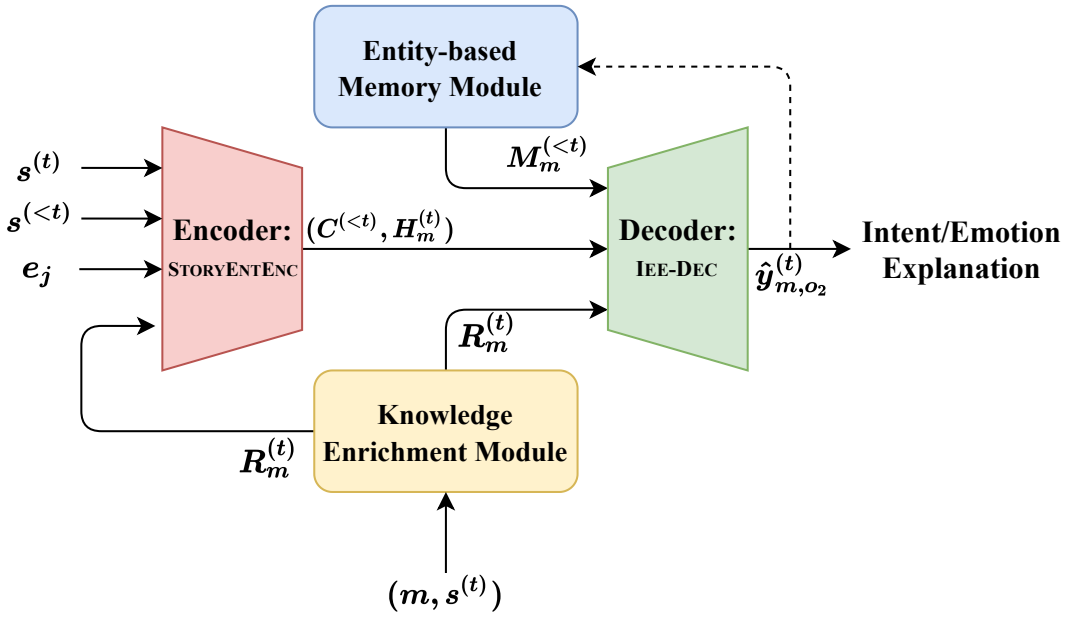


FIGURE 5.2: Overview of our NEMO model.

- *Knowledge-Enrichment Module* (KEM): Following a recent work by [VR21a], we utilize a pretrained EVENTBERT for this component. EVENTBERT leverages social commonsense knowledge to sharpen the social event embeddings with semantic and pragmatic attributes. Here, the pragmatic properties refer to the human's inferred implicit understanding of event actors' intents and feelings or reactions. We feed the mental state attribute m and the current story sentence $s^{(t)}$ as our input and get a sentence-level attribute-specific pragmatics-aware embedding $R_m^{(t)}$ as the output of this module.

- *Story Entity Encoder*(STORYENTENC): Our modified Transformer-based encoder is employed to produce prior story context embedding $(C^{(<t)})$ and entity-aware representation (H_m^t) of the current story sentence $(s^{(t)})$

consolidating the prior story sentences ($s^{(<t)}$), entity (e_j) and mental state attribute-specific pragmatics-aware knowledge embedding ($R_m^{(t)}$) obtained from (KEM).

- *Entity-based Memory Module* (EMM) : This module is used to dynamically track the prior mental states of characters in the narrative so that the generated explanations are coherent to the previous events in the narrative. Therefore, we keep track of previously generated mental state representations in a separate memory indexed using each entity (e_j) and mental state attribute information (m) and denoted as $\mathcal{M}[e_j, m]$. This module is accessed during the decoding phase by attending over memory cells to obtain attribute-specific prior mental state embeddings ($M_m^{(<t)}$).

- *Intent-Emotion Explanation Generator*(IEE-DEC): Our two pass-iterative decoder generates intent and emotion explanations by processing the encoder outputs ($C^{(<t)}, H_m^{(<t)}$), KEM output ($R_m^{(t)}$), and attribute-specific entity e_j's prior mental state embeddings $M_m^{(<t)}$retrieved from EMM.

5.4.1 Story Entity Encoder

Figure 5.3 presents a closer look into the model architecture. A variant of the conventional Transformer encoder is used to produce an entity-aware representation of the story. We introduce additional sub-layers to incorporate prior context, entity, and mental state attribute information. Our encoding strategy, STORYENTENC$(\cdot)$, is defined as:

$$(C^{(<t)}, H_m^{(t)}) = \text{STORYENTENC}(s^{(t)}, s^{(<t)}, e_j, m) \tag{5.1}$$

where $e_j \in \mathcal{E}$ is the entity under consideration, $H_m^{(t)}$ is the resulting entity-aware representation of the story at t^{th}-step. Inspired by [Her+15], we identify character names in a story using Coref systems and replace them with abstract markers to prevent degenerate solutions. We do not replace words related to social roles. We randomly permute these entity markers from a set of generic markers reused across multiple stories to primarily distinguish them from other entities in the story during our training and testing process. This allows us to embed unseen entities in new stories. We denote the story-specific entity embeddings as $E_{e_j} \in \mathcal{R}^{d_e}$.

Our STORYENTENC is composed of a stack of N_s identical layers. To create an entity-specific understanding of the story, we perform the following steps: (a) concatenate the character information along with the current sentence to produce entity or character-aware representation of the story sentence, (b) introduce an additional context-attention sub-layer that integrates story context into the encoder, and (c) fuse knowledge representation related to specific mental state attributes. The entity concatenated input sentence is given as: $[CLS]\ e_j\ [SEP]\ w_1^{(t)}, ..., w_L^{(t)}\ [SEP]$ and $E_s^{(t)}$ is its corresponding matrix containing d_w-dimensional word-embedding vectors (in our case, $d_e = d_w$). Using steps (a) and (b), we integrate the interactions between entity-specific information from the current story sentence and its prior context. This process is given as follows:

$$U^{(l)} = \text{MHA}(H_s^{(l-1)}, H_s^{(l-1)}, H_s^{(l-1)}) \tag{5.2}$$

$$V^{(l)} = \text{MHA}(U^{(l)}, C^{(<t)}, C^{(<t)}) \tag{5.3}$$

$$H_s^{(l)} = \text{FFL}(V^{(l)}) \tag{5.4}$$

where l is the encoding layer, $l \in \{1, 2, ..., N_s\}$ and $H_s^{(0)} = E_s^{(t)}$, $C^{(<t)}$ is the prior story context embedding as computed in Section 5.4.1 and $H_s^{(l)}$ is the

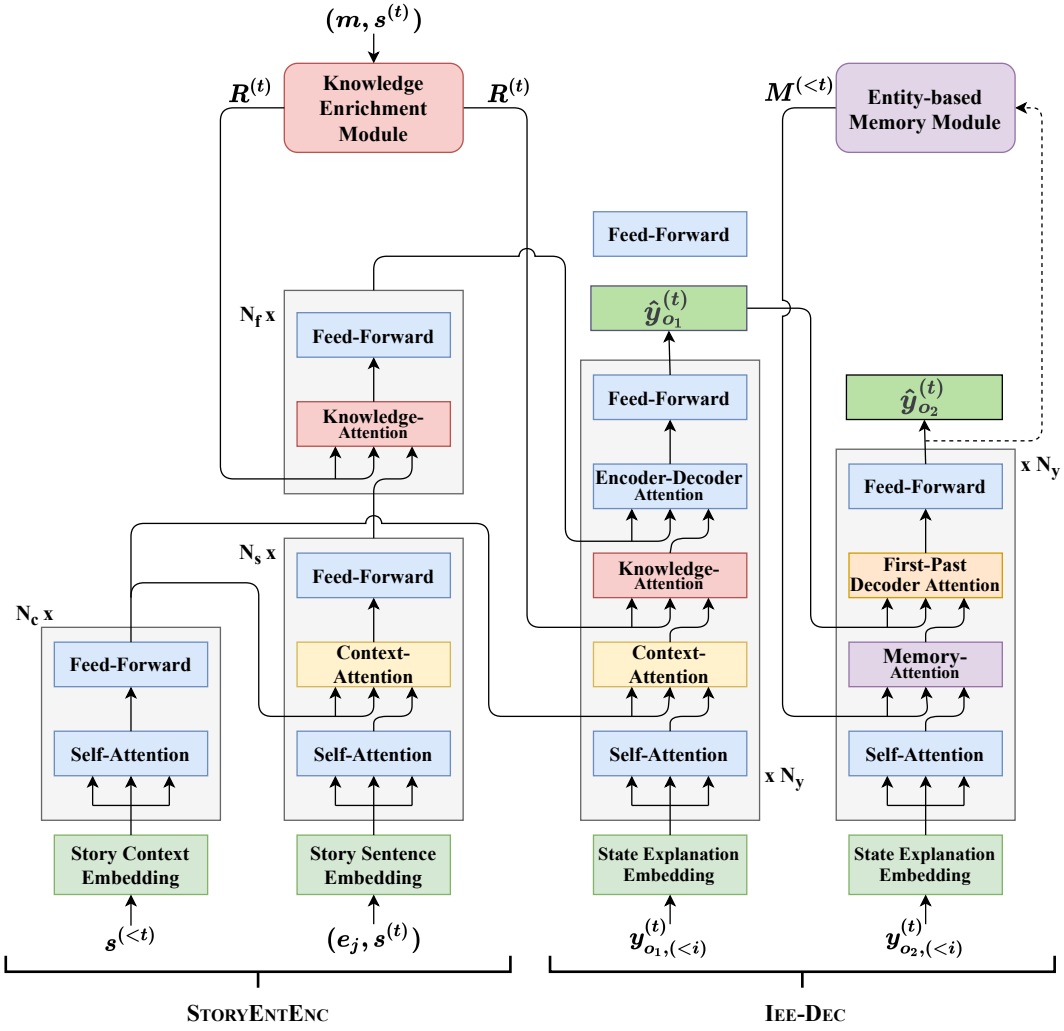


FIGURE 5.3: Illustration of the full architecture of our NEMO model.

embedding of the source sentence at the l^{th} layer. Finally, we fuse the knowledge representation $(R_m^{(t)})$ related to specific mental state attribute (m) obtained from (KEM) with the output at N_s^{th} layer $(H_s^{(N_s)})$. The fusion step involves N_f additional Transformer layers with the context-attention replaced by knowledge-attention i.e. $\text{MHA}(H_s^{N_s}, R_m(t), R_m(t))$. We found in preliminary experiments that even a single fusion layer is effective in outperforming our baselines. The output from the fusion layer is the final encoded story representation, $H_m^{(t)}$, encapsulating context, entity and attribute-specific knowledge information.

Context-Attention & Gating

We implement standard Transformer encoder layers for computing the story context information from previous sentences $s^{(<t)}$. For the prior story context $s^{(<t)}$, we insert a $[CLS]$ and $[SEP]$ token at the start and end of each sentence, respectively. Since we add new sub-layers in this work, we introduce a gating mechanism instead of residual connections to prevent the uncontrolled influence of information from sub-layers over the current sentence representation:

$$\beta = \sigma(W_1 H + W_2 f(H)) \tag{5.5}$$

$$G(H) = \beta \odot f(H) + (1 - \beta) \odot H \tag{5.6}$$

where f refers to the sub-layers, $\sigma(\cdot)$ is a sigmoid function, W_1, W_2 are learnable parameters.

5.4.2 Intent-Emotion Explanation Generator

Motivated by human cognitive behaviors, we explore the process of deliberation into the sequence generation framework [Xia+17]. This is implemented as a two pass-iterative decoding strategy. During the first pass, the decoder generates a rough draft of the explanations ($\hat{y}^{(t)}_{o_1}$) by considering sentence-level knowledge along with encoder outputs and prior context. The first step decoding outputs are fed to the second pass decoder along with entity's mental state context obtained from an entity-based memory module EMM. Formally, the two-step decoding procedure is denoted as:

$$\hat{y}^{(t)}_{m,o_2} = \text{IEE-DEC}(H^{(t)}_m, C^{(<t)}, R^{(t)}_m, M^{(<t)}_m) \tag{5.7}$$

where $R^{(t)}$ is the sentence-level knowledge embedding from KEM, $M^{(<t)}_m$ is the attribute-specific entity's prior mental state embeddings retrieved from

EMM. In order to generate entities' intent and emotion explanations, we introduce artificial tokens associated with mental state attributes as the start token. These special tokens could be one of the following mental state attributes, $m \in \{xIntent, xReact\}$. For brevity, we drop the subscript m from the equations.

First-pass Decoding

Just like the encoder, our decoder has N_y stacked identical layers. We augment each layer with context and knowledge-attention sub-layers. While the former provides prior story context representation (extracted from $s^{(<t)}$), the latter captures the attribute-specific sentence-level knowledge information ($R^{(t)}$). The first-pass decoding procedure is explained as follows:

$$U'^{(l)} = \text{MHA}(H_{o_1}^{(l-1)}, H_{o_1}^{(l-1)}, H_{o_1}^{(l-1)}) \tag{5.8}$$

$$V'^{(l)} = \text{MHA}(U'^{(l)}, C^{(<t)}, C^{(<t)}) \tag{5.9}$$

$$W^{(l)} = \text{MHA}(V'^{(l)}, R^{(t)}, R^{(t)}) \tag{5.10}$$

$$Z^{(l)} = \text{MHA}(W^{(l)}, H^{(t)}, H^{(t)}) \tag{5.11}$$

$$H_{o_1}^{(t)} = \text{FFL}(Z^{(l)}) \tag{5.12}$$

where $l \in \{1, 2, .., N_y\}$, $H_{o_1}^{(l-1)}$ is the output from previous layer, and $H_{o_1}^{(0)} = [y_0^{(t)}, y_1^{(t)}, ..., y_{i-1}^{(t)}]$ denotes the representation of words generated up until the i^{th} step ($y_{<i}^{(t)}$). Before feeding our computed representations to a feed-forward layer, we integrate the representation of the current sentence from the encoder using encoder-decoder attention. At the end of N_y layers, we compute word probabilities for the first-pass decoded sequence: $P(\hat{y}_{o_1}^{(t)}) = softmax(H_{o_1}^{(N_y)})$. Here $\hat{y}_{o_1}^{(t)}$ is the first-pass decoding output.

Second-Pass Decoder

During the second-pass decoder, we contextualize the current entity states' using entity's prior mental state embeddings ($M^{(<t)}$) stored in an entity-specific external memory (EMM) in combination with the first-pass decoder outputs:

$$W'^{(l)} = \text{MHA}(U''^{(l)}, M^{(<t)}, M^{(<t)}) \tag{5.13}$$

$$Z'^{(l)} = \text{MHA}(W'^{(l)}, \hat{H}_{o_1}^{(t)}, \hat{H}_{o_1}^{(t)}) \tag{5.14}$$

where $U''^{(l)}$ is the second-pass counterpart of self-attention sub-layer ($U'^{(l)}$) and $\hat{H}_{o_1}^{(t)}$ is the representation of words generated during the first pass. The polished mental state explanations are computed as: $P(\hat{y}_{o_2}^{(t)}) = softmax(H_{o_2}^{(N_y)})$, where $H_{o_2}^{(l)} = \text{FFL}(Z'^{(l)})$ is the feed-forward sub-layer output. Thus, $\hat{y}_{o_2}^{(t)}$ is the polished decoded output.

5.4.3 Knowledge-Enrichment Module

Knowledge-Enrichment Module (KEM) can be viewed akin to a semantic memory [BC08]. Generally, semantic memory refers to a long-term storehouse of general knowledge related to events, facts, and concepts. The core idea is to encode a story sentence into a pragmatics-aware embedding. The pragmatic components refer to the implied emotions and intents associated with the events in the story text. By leveraging social commonsense knowledge explained in Section 3.1.2, we follow a recent work of [VR21a] and utilize the EVENTBERT as our KEM to pretrain and effectively embed both semantic and pragmatic aspects of social events.

The input is a concatenation of mental state attribute $m \in \{xIntent, xReact\}$ with the story sentence $s^{(t)}$. This is fed through the EVENTBERT to produce attribute-specific contextualized social event embeddings, $R_m^{(t)}$. This

encoding step is followed by an attentive pooling function that attends over contextual embeddings to output a summarized pragmatics-aware embedding $r_m \in \mathcal{R}^{d_h}$ reflecting intents ($m = xIntent$) and emotional reactions ($m = xReact$). We learn these representations by pretraining using an N-pair loss (as in [VR21a]) for each intent or emotion explanations. By training on data from social commonsense knowledge sources, we enable the model to learn pragmatics-aware representation of the social events. While the contextual vectors $R^{(t)}$ are used during encoding and decoding phases, the summarized vectors $r_m^{(t)}$ are used to initialize our entity-based memory module (EMM). MOre details on this component has already been described in Chapter 4.

5.4.4 Entity-based Memory Module

Entity-based Memory Module can be seen as an episodic memory that ideally stores the mental states of characters in a specific narrative. To track entity-specific mental state representations, we utilize a memory, $\mathcal{M}$, containing separate memory cells for each entity e_j and mental state attribute m. Therefore, memory is indexed using entity embeddings (E_{e_j}) and mental state attribute embeddings (E_m)). For simplicity, we denote it as: $\mathcal{M}[e_j, m]$. The memory operations are explained as follows: $\mathcal{M} = (\mathcal{K}, \mathcal{A}, \mathcal{V})$, where key $\mathcal{K}$ is tied with entity embeddings, $\mathcal{A}$ refers to the mental state attribute m and $\mathcal{V}$ contains the attribute-specific target-size context vectors.

Memory Attention: Our decoder applies a multi-head attention mechanism over prior mental state representations of an entity $M_m^{(<t)}$ for each mental state attribute m. For a specific entity e_j and mental state attribute m, we retrieve $(t-1)$ memory cells from $\mathcal{M}[e_j, m]$ by masking the future time steps. Finally, we inject the sequence-order information using positional encoding [Vas+17] to get $M^{(<t)}$ (drop the subscript m to be consistent with previous

notations).

Memory Write: We keep track of prior mental states by storing their representations in our memory $\mathcal{M}[e_j, m]$. It is possible to limit the memory allocated to each entity for prior context (say n-previous sentences). However, we don't set such limits in this work. We initialize the memory with sentence-level pragmatics-aware summarized vector, $r_m^{(t)}$. For the write operation, we apply a gating mechanism to store the final decoder hidden state of $\hat{y}_{o_2}^{(t)}$ given as $h_{\hat{y}_{o_2}}^{(t,L)}$ at the t^{th} memory cell:

$$\gamma = \sigma(W_r r_m^{(t)} + W_h h_{\hat{y}_{o_2}}^{(t,L)}) \tag{5.15}$$

$$\mathcal{M}[e_j, m, t] = \gamma \odot r_m^{(t)} + (1 - \gamma) \odot h_{\hat{y}_{o_2}}^{(t,L)} \tag{5.16}$$

where W_r and W_h are learnable parameters. In our experiments, we find that this method is simple yet effective.

5.5 Training & Hyperparameters

Our aggregated data is split into train, validation, and test sets at 70-10-20 split. Following [Xio+19]'s work, our model is trained to minimize the negative log-likelihood of predicting each word during both the decoding steps: $\mathcal{L} = \mathcal{L}_{mle1} + \mathcal{L}_{mle2}$. To handle our weakly-annotated data, we perform phase-wise training of our model. We pretrain our model using all the social commonsense knowledge data where the entity or character information is concatenated with the input text during the first phase. The memory cells are initialized to zero, and the model learns to produce sentence-level explanations. The second phase involves modeling the current story sentence along with the prior narrative context. We initialize the memory with pretrained sentence-level knowledge embedding $r_m^{(t)}$ once for a mini-batch and further

update them with noisy explanations. This exposes the model to its potential test-time errors and guides the model to learn robust parameters.

Using grid-search, we tune the hyperparameters and the best configuration ($N_c = 2, N_f = 1, N_s = 12, d_h = 768$ and 12 attention heads) is obtained based on validation set perplexity. To prevent overfitting, we use dropout with a rate of 0.2. By default, we experiment with GloVe vectors and ELMo-based contextualized embeddings (usually mentioned during evaluation). We use Adam as our optimizer with a learning rate of $\alpha = 0.0002$ [KB14] and a training batch size of 8. We use greedy decoding at training time, but utilize beam-search with a beam size of $k = \{3, 5, 10, 12\}$ [BCB14; SHB15] at inference time.

5.6 Experiments

In this section, we describe the various evaluation settings: datasets, baselines, model variants, modes, and metrics. We designed our experiments to study the following research questions:

RQ1: How well does our model perform compared to other baselines in the explanation generation task? How much does each component impact the overall performance?

RQ2: Can our model representations be used to perform state classification based on labeled motivation and emotional reaction categories?

RQ3: Do the learned mental state representations exhibit transfer capability to a downstream task?

Dataset	#Stories	#Motives	#Emotions
Personal Narratives	300	882	1418
StoryCommonsense	2500	6831	13785

Table 5.1: Test set statistics for explanation generation task: This includes number of annotated stories and number of character-lines with motives and emotions.

5.6.1 Explanation Generation Task (RQ1)

Dataset

We run experiments on (a) the manually annotated gold explanations for sampled data from personal narratives corpus and (b) the benchmark character psychology dataset – StoryCommonsense [Ras+18c]. Table 5.1 summarizes the dataset used for evaluation of our explanation generation task.

Baselines

We compare our model's performance to different baseline methods. We follow a model architecture for the baseline methods as in [Ras+18c], where they compute an encoded vector by concatenating the current sentence representation along with the entity-specific context (involving sentences where a particular entity appears). These methods are enlisted as follows:

- **LSTM** [Ser+17], which is a hierarchical RNN-based encoder-decoder model. The sentence tokens are encoded using a bi-LSTM. The entity-specific vector, computed using a similar method, is then concatenated with the sentence vector.

- **Ren** [Hen+16], which is a recurrent entity network updating entity states in a dynamic long-term memory. A memory cell is initialized for every entity in the story and updated after reading every sentence. The memory vector in the cell corresponding to the entity under consideration is the final encoded vector.

- **NPN** [Bos+17], which performs dynamic entity tracking by explicitly modeling actions as state transformers. Memory is initialized and accessed as in REN.

- **GPT** [Rad+18], which is a fine-tuned transformer-based language model architecture. The input setup consists of the concatenation of entity marker, story context tokens, current sentence tokens, and mental state attribute token m separated by special $[SEP]$ tokens. This is closely related to how GPT is used in [Bos+19].

Model Variants

By evaluating on the personal narrative corpus, we assess the impact of three of our model components: KEM (semantic memory), EMM (episodic memory), IEE-DEC (deliberation decoder). To comprehensively study their impact, we remove them one at a time as model variants and evaluate their impact on the performance in the explanation generation task.

Metrics

Due to the short sequence length of generated explanations and the high possibility of producing similar explanations in multiple ways, we avoid word-overlap based metrics and instead compute embedding-based metrics such as embedding average and vector extrema for evaluating explanation generation quality [Liu+16; HKN18]. Embedding average calculates sentence-level embeddings by averaging the word embeddings of each token in a sentence. Vector extrema metric takes the most extreme value for each dimension amongst all word vectors in the sentence and uses that value in the sentence-level embedding. To compare the ground truth and generated explanation, we compute the cosine similarity between their respective sentence-level vectors. These metrics Additionally, these metrics are useful

Models	Motivation		Emotion	
	Avg	VE	Avg	VE
HRED	58.43	48.78	52.05	51.18
REN	58.96	49.87	53.59	52.14
NPN	59.03	50.02	52.63	51.76
GPT	63.56	54.77	56.74	56.09
NEMO	**69.27**	**59.78**	**62.88**	**61.34**
COMET repl.	66.55	58.23	60.78	60.21
w/o EMM	67.16	57.64	59.74	58.38
w/o KEM	65.38	56.86	60.58	60.06
w/o IEE-DEC	66.92	58.17	60.42	59.83

(A) Personal Narrative Corpus

Models	Motivation		Emotion	
	Avg	VE	Avg	VE
Random	56.02	45.75	40.23	39.98
LSTM	58.48	51.07	52.47	52.30
REN	58.83	51.79	53.95	53.79
NPN	57.77	51.77	54.02	53.85
GPT	60.19	52.95	55.68	55.47
NEMO	**66.25**	**59.16**	**62.78**	**61.92**

(B) STORYCOMMONSENSE dataset

TABLE 5.2: Automatic evaluation results on (a) Personal Narratives corpus & (b) STORYCOMMONSENSE dataset. Bold face indicates leading results for the corresponding metric.

for comparison with the previous benchmark used for the generation task [Ras+18c].

Results

The main results of our evaluation on Personal narratives and STORYCOMMONSENSE datasets are summarized in Table 5.2a and Figure 5.2b respectively. We observe that our complete model achieves an absolute mean improvement of $\sim$ 9% and $\sim$ 12% over a fine-tuned GPT model using the embedding average metric of the generated intent and emotion explanations respectively across both the datasets.

Effect of architectural choices By training variants of our NEMO model with and without specific components of the model, we are able to ascertain their importance for the task at hand. Table 5.2a shows that our KEM and EMM yield significant boost to the overall performance. The dip in performance on intent generation is more pronounced when KEM is removed while EMM is critical for the improved performance of emotion generation. We intuit the reason to be the additional sentence-level commonsense knowledge infused by KEM leading to better generations of intents while entities' prior states from EMM guiding the overall prediction of the current emotional state.

Effect of Knowledge Embeddings From Table 5.2a, it is clear that KEM provides really good performance gains. Further, we replace the knowledge embeddings obtained from KEM with the embeddings extracted from COMET [Bos+19]. COMET is a framework that adapts the language model weights to produce diverse commonsense knowledge tuples. The scores reported in Table 5.2a indicate that there is a significant advantage of using KEM embeddings over COMET. Also, we note that COMET only provides a small marginal improvement in comparison to a NEMO model without KEM component. But the addition of our KEM component provides a huge jump in performance, specifically while generating motives. This can be attributed to the social role information, a characteristic of our social commonsense knowledge resource, utilized by our KEM module. We verify this in the error analysis (see Section 5.6.2).

Human Evaluation of Trajectories

We conduct a human evaluation to test the effectiveness of our NEMO model in generating motivation and emotion explanation trajectories. Our experiment compares our model explanations to those obtained from GPT-based

model. We randomly select 100 stories and present the story, character, and the visualization of trajectories to three workers. The workers then select the trajectory that best matches the characters' mental states. The inter-annotator agreement had a Fleiss' $\kappa = 0.74$ and $\kappa = 0.78$ for intent and emotion trajectories respectively, indicating substantial agreement among the workers. Moreover, 47 intent and 56 emotion trajectories had a unanimous agreement among three workers, of which 45 intent and 52 emotion trajectories were in favor of trajectories generated by NEMO. Based on the majority agreement, the workers selected our intent and emotion trajectories for 81% and 83% of the presented stories, respectively. Thus, it is clear that our model is able to generate better explanation trajectories.

Qualitative Analysis

Effect of Context We investigate the effect of context in producing convincing explanations for our text by filtering null attention and plotting an attention map between context and source text (see Figure 5.4). Notably, this particular attention head (head-6) maps specific source words to their relevant context words. The attention head's focuses on the following words: "relocate" $\mapsto$ {"lived","beach", "hurricane"} and "they" $\mapsto$ {"jennifer","her", "family"}. Further, we also show sample generation with and without the context information. It is evident from these examples that NEMO can identify particular aspects of the context that are relevant (e.g., antecedents, spatial concepts) and leverage them to produce appropriate explanations.

Effect of Two-pass decoding Step Figure 5.5 provides sample motivations generated by our proposed model in multiple passes along with GPT (as it performs competitively for our task). We demonstrate our model's ability to generate explanations from the narrator's perspective (1^{st} person) and that

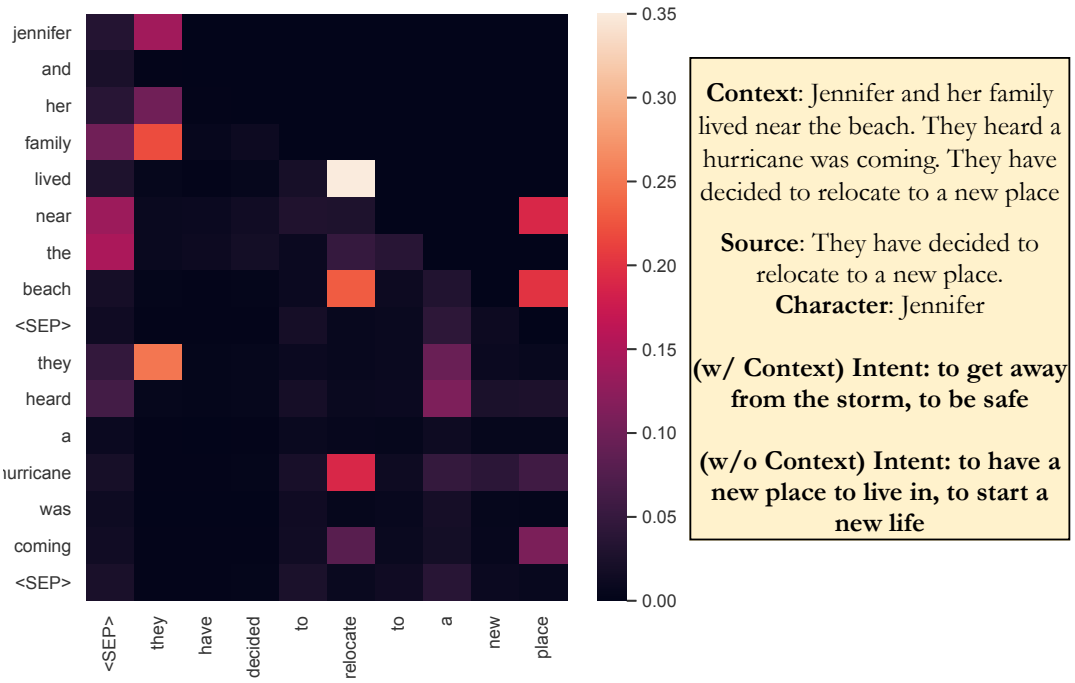


FIGURE 5.4: Attention map (head-6) between context and source. On the x-axis are the source tokens, on the y-axis the context tokens.

of another entity/character in the narrative. The result also shows improvement after the second-pass decoding.

Input Text	**Context:** I loved Mary intensely. But she wanted to be only friends with me. **Source:** She found a guy called John	**Context**: I was joining a grad school at time when ethnic problems were rife. **Source**: As I was crossing the road to get to my class, there was a girl who was resisting the police brutality, leading from front and later I joined it too.
Models	**Entity/Character: Mary**	**Entity/Character: I (myself)**
GPT	• none • to be loved • to be happy	• to get to the class • to gain knowledge • to be safe
Our Model (First pass)	• to have a relationship • to be with him • to be loved by someone	• to get to the class quickly • to get to the class in problems • to defend someone
Our Model (Second Pass)	• to have a relationship **with john** • to **be friends with john** • to **have a relationship with the guy**	• to get to the class **on time** • to **stand up for a cause** • to **defend someone in a situation**

FIGURE 5.5: Generation of motivation explanations in multiple decoding steps.

5.6.2 State Classification Task (RQ2)

Dataset

The STORYCOMMONSENSE dataset comprises over 300k low-level annotations for motivations and emotions across 15,000 short stories selected from ROCStories training set [Ras+18c]. This dataset includes the categorization of motivations and emotional reactions based on different classical theories of psychology.

Experimental Settings & Baselines

We conduct experiments under the following settings:

- **Zero-shot (ZS)** In this setting, we map the generated emotion explanation to one of the 8 Plutchik's categories via nearest neighbor search in the word-embedding space: $\bar{y}_c = argmax_{c \in C}(cos(E_{\hat{y}_{xReact}}, E_c))$, where $c \in C$ is the label related to Plutchik's categories. Without any further fine-tuning, we compare our results against COMET-CGA [BC19] and use their word formulation setup for labels.

- **Supervised (SS)** We fine-tune our trained model using a feed-forward layer on the top of the encoder output. Additionally, we experiment with (NEMO$_E$) and without (NEMO$_{NE}$) annotated explanation training. In addition to the baselines in the original work, we compare against – BiLSTM + Self-Attention (**BM**) and BiLSTM + Self-Attention + Knowledge (**BM+K**) which incorporate multihop knowledge paths using graph-based algorithms [PF19] for predicting human needs (motivation categories). Additionally, we report scores from a recent work [Gao+20] that uses label semantics (referred to as LS) and track label-label correlation for emotion inference task.

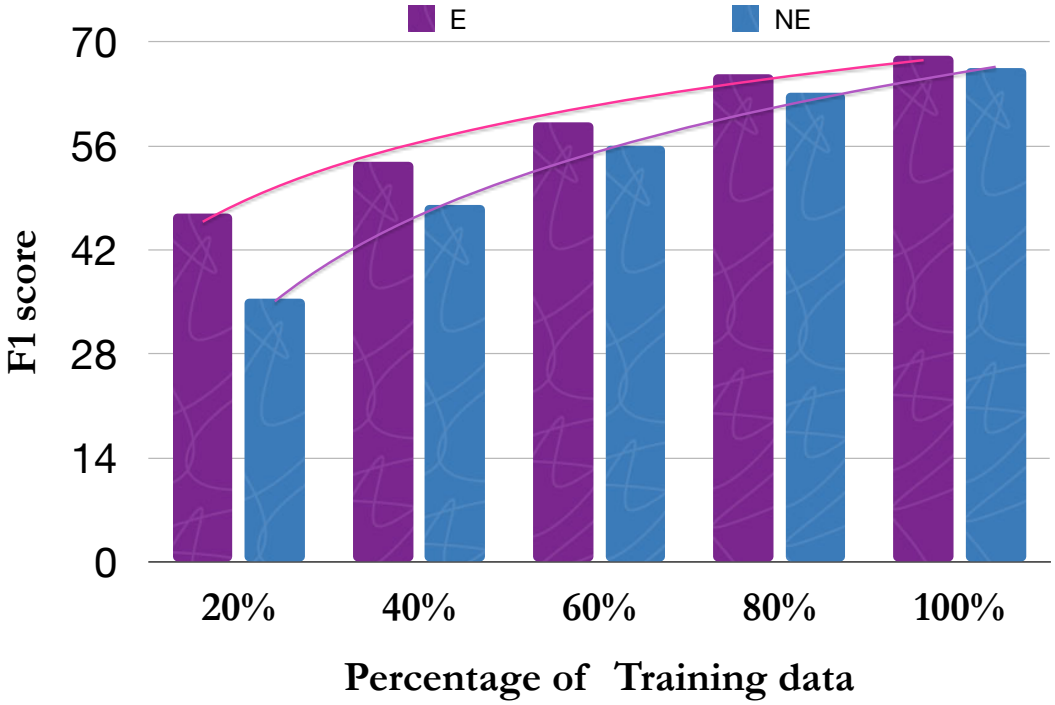


FIGURE 5.6: Prediction performance under low-resource (LR) settings (limited amounts of training data).

- **Low-resource regimes (LR)** This scenario has a significant practical interest, specifically in adapting our model to domains with a small amount of in-domain labeled data. Having trained on personal narratives corpus, we simulate low resource regimes by varying the percentage of training examples from STORYCOMMONSENSE state classification dataset.

Metrics

Consistent with prior study [Ras+18c], we compute the micro-averaged F_1 scores for the state classification task: Maslow, Reiss and Plutchik states.

Results

Visibly, our models (in Table 5.3) outperform the state-of-the-art methods significantly in both settings. With ELMo-based embeddings, the improvements are even more pronounced. For the zero-shot settings, we report the scores

Models	Maslow	Plutchik	Reiss
COMET-CGA (ZS)	_	19.30	_
COMET-CGA (T)	_	27.50	_
NEMO (ZS)	_	42.61	_
LSTM	34.55	28.81	24.51
CNN	35.23	30.04	24.21
REN	33.57	30.15	20.53
NPN	31.69	30.29	17.75
BM	53.54	_	26.57
BM+K	56.69	_	32.96
BM (ELMo)	59.81	_	35.49
BM+K (ELMo)	61.72	_	36.70
LS	_	65.88	_
NEMO$_E$	69.77	68.16	45.76
NEMO$_{NE}$	67.37	66.57	46.21
NEMO$_E$ (ELMo)	**72.09**	**71.26**	**48.52**
w/o EMM	69.13	67.19	45.50
w/o KEM	66.28	68.61	43.92
COMET repl.	66.95	69.14	43.92

TABLE 5.3: State classification performance under supervised settings. ZS: Zero-Shot Settings, T-Tuned Hyperparameters as reported in [BC19]

directly from the original work [BC19]. We find a similar pattern in the state classification task for the supervised setting as in Section 5.6.1. The impact of COMET is only marginally felt while the KEM component provides a relatively huge performance boost. Interestingly, our results in Figure 5.6 suggest that the model variant fine-tuned with explanations learns faster with lesser in-domain labeled data than its counterpart without explanation fine-tuning. We note that both these models outperform several baselines with less than 40% of training examples. We believe that the explanation fine-tuning further sharpens the learned mental state representations as the annotations are much cleaner than our aggregated personal narratives corpus.

Error Analysis

Decoding Phase During the beam-decoding step for sentences irrelevant for a particular entity, "none" can only be predicted once, which causes other candidates in the beam to be incorrect if "none" is the appropriate answer. However, we posit that the embeddings hold richer information than the explanations generated due to the limitations in the way we implement the decoding. For the explanations, we observe that the models miss out on quantities that are expressed as numbers or ambiguous phrases used in social media personal narratives. Figure 5.7 shows an example where a small replacement to the input produces better explanations.

Emotion State Classification A noticeable trend in the categorization task is the high level of cross-predictions among related emotions. Several misclassifications occur between joy-surprise and anger-disgust categories. The subtle difference between those emotion pairs makes it harder for the models to distinguish them in some cases clearly.

Intent State Classification Since we observed only a marginal improvement with the addition of the COMET module, we compare the difference in errors made by COMET replaced NEMO model in comparison to our complete NEMO model for predicting Maslow's motivation categories. We gauge that COMET replaced model made more errors (in $\sim$ 24.5% of the cases) when the stories contain more than one social role information. This validates our claim that the social role information captured by our NEMO with KEM module is beneficial for both the classification and generation task.

Sample Generation
Context: Since I was a kid, I wanted to be a writer. I wanted to make people happy and have them be super excited to read what I did. Well last night, this became a reality and I got another book published. I was proud, and happy. I was excited to share with friends. But **only four** friends cared to even read it. **Source:** I'm legitimately heartbroken and disinterred
Entity/Character: I (myself)
• to express my feelings • to express my feelings about the book • to be a great author
Replace "only four" with "only a handful of friends"
• to express my feelings • to have a lot of friends to read it • to have a lot of friends to read the book

FIGURE 5.7: Sample generations showcasing the limitations of NEMO.

5.6.3 Application: Empathetic Dialogue Generation (RQ3)

Natural social interactions require humans to recognize and infer others' implied emotions and respond appropriately by acknowledging their underlying feelings. Since NEMO infers motivations and emotion states from stories, we posit that the embeddings learned from such a model can lead to improved performance on this dialogue generation task.

Models	PPL	AVG BLEU
Fine-Tuned	21.24	6.27
Fine-Tuned Large	**16.55**	**8.06**
EmoPrepend-1	24.30	4.36
TopicPrepend-1	25.40	4.17
Ensem-DM	19.05	6.83
Ensem-SCS+	17.06	7.64

TABLE 5.4: Automatic evaluation metrics on ED test set. Ensem-SCS+: model incorporating our learned embeddings.

Dataset

We use EMPATHETICDIALOGUES (ED) dataset, introduced by [Ras+18d], for evaluating the ability of NEMO representations to improve generation of empathetic responses. The dataset consists of 25k personal dialogues grounded in specific emotional situations where a speaker was feeling a given emotion, with a listener responding. The train/ val/ test split was 19533/ 2770/ 2547 conversations, respectively.

Model & Baselines

Following Rashkin et al.'s prior work [Ras+18b], we experiment with the ensemble of encoders that augments the encoders to incorporate the embeddings extracted from pretrained architectures. The ensemble model that incorporates our mental state representations is referred to as "Ensem-SCS+". We compare our ensemble model with other well-performing benchmarks reported in [Ras+18b] involving pretrained external predictors:

- **Ensem-DM**: An ensemble model with supervision from trained Deepmoji system [Fel+17].
- **EmoPrepend-1**: Add an emotion label to the beginning of the token sequence as encoder input. This is obtained from a separate classifier that predicts emotion labels from the description of the situation.

- **TopicPrepend-1**: Similarly, the top predicted label from the supervised topic classifier is merely prepended to the beginning of the token sequence as encoder input.

Results

For the above baselines, we report the values directly from the original work. Table 5.4 shows that our Ensem-SCS+ model produces significant improvement in automated metrics, quantifying the impact of using our learned representations. Our model with a relatively lower number of parameters is able to perform closer to the best performing large model.

5.7 Conclusion

In this paper, we present a Transformer-based method to model the mental states of characters related to the events in the personal narratives. Using data mining and information extraction techniques, we aggregate weakly-annotated data to train our model known as NEMO. We show that the proposed method is able to outperform several baselines in mental state tracking task. We also observe that the pretraining on weakly-annotated data helps in improving the overall performance under low-resource settings. We believe that further improvements can be achieved in explanation generation and state categorization in cases where the text contains character-irrelevant content or non-events by introducing specialized knowledge in our model. Our analysis also demonstrated the transferability of our learned representation in a downstream empathetic response generation task. Future work could investigate the applicability of these mental state representations in modeling vital elements of narrative structures.

Chapter 6

Modeling Narrative Structure in Short Personal Narratives

6.1 Introduction

Narratives are the fundamental means by which people organize, understand, and explain their experiences in the world around them. Researchers in the field of psychology maintain that the default mode of human cognition is a narrative mode [Bec15]. Humans share their personal experiences by picking specific events or facts and weave them together to make meaning. These are referred to as personal narratives, a form of autobiographical storytelling that gives shape to experiences. [Pol88] suggested that personal narratives, like other stories, follow broad characteristics involving: (a) typically a beginning, middle, and end, (b) specific plots with different characters and settings, or events. Often, characters learn something or change as a result of the situation or a conflict and resolution, but not always. Some of these characteristics provide a basis for the organizational framework of a story, commonly referred to as the narrative structure or the storyline. The growing amount of personal narrative text information in the form of social media

posts, comments, life stories, or blog posts presents new challenges in keeping track of the storyline or events that form the defining moments of the narrative. Several recent works [Dor+18; Yua+17; CLG17; Ko+18; Mos+17] have made efforts to advance the research in narrative comprehension. However, the development of computational models that automatically detect and interpret different structural elements of a narrative remains an open problem. Discovery of structural elements of a narrative has many applications in: (a) retrieval of narratives based on similar dramatic events or concepts instead of keywords [MAP91; FW06; MLL91], (b) linking related stories that form a narrative thread towards theme generation [BS13; Sun07], (c) summarization of stories [Leh81; Pap+20] and (d) story ending prediction or generation [CCY19; Li+13; Mos+17], (e) commonsense reasoning [Goo+12; GBS11], to list a few.

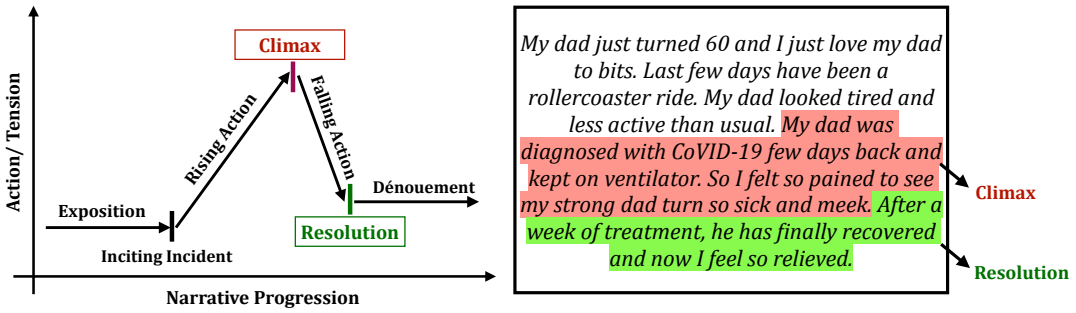


FIGURE 6.1: (Left) Freytag's Pyramid. (Right) Highlights of climax and resolution for a sample personal narrative.

Researchers in narratology have analyzed various components of a narrative that contribute to a notion of plot quality referred to as 'tellability'. It is commonly derived from certain structural properties used in narrative theory. Several narrative theories have been proposed such as Freytag [Fre94], Prince [Pri12], Bruner [Bru91b; Bru09], Labov & Waletzky [LW97], to name a few. These theories explain different elements of a narrative structure containing typical orderings between them. Certain elements of the narrative structure are correlated across different narrative theories. For example, Bruner's

'breach in canonicity' [Bru91b] could correspond to (a) Freytags 'climax' – referring to the 'turning point' of the fortunes of the protagonist [AH14] or (b) Labov's 'most reportable event' – describing the event that has the greatest effect upon the goals, motivations and emotions of the characters (participants) in the narrative [LW97; Lab06]. Shorter narratives tend to consist mostly of complicating actions that culminate in the MRE or climax and instances of events that reach a 'resolution' stage indicated by a swift drop in dramatic tension. In comparison, the other structural elements are more likely to occur in longer narratives. Figure 6.1 (Left) shows the Freytag's pyramid containing the key elements of the narrative structure, and Figure 6.1 (right) includes highlights of climax (used interchangeably as MRE) and resolution for a sample personal narrative. Thus, our work aims to leverage computational approaches at the intersection of information retrieval, NLP, and psychological aspects and model the key elements of narrative structure – MRE and resolution. As a working definition, we consider an MRE/climax to be contained in the sentence(s) based on the following criteria: (a) it is an explicit event at the highest tension point of the story, and (b) it is the only event that can be reported as the summary of the story. Similarly, an event qualifies as 'resolution' if it usually occurs after the MRE and resolves the dramatic tension in the narrative.

Papalampidi et al. [PKL19] introduced a dataset consisting of movie screenplays and plot synopsis annotated with turning points. Few attempts have been made at annotating elements of high-level narrative structures [Li+17] and automatically extracting them from the text. Ouyang et al. [OM15]'s study on predicting MRE in narratives is the closest work to the problem considered in this study. While most of these methods rely on syntactic, semantic, surface-level affect or narrative features obtained using hand-engineering or pretrained semantic embedding methods to model narrative structure, we investigate the role of the protagonist's psychological states in capturing the

pivotal events in the narrative and their relative importance in identifying the elements of narrative structure – Climax and Resolution. We find the basis for this study in prior theoretical frameworks [Mur03; Rya86; OM14; Leh81; Sch16] that emphasize (a) how narrative structure organizes the use of psychological concepts (e.g., intentions, desires, and emotions) and mediates all the human interactions and their social behavior, and (b) how protagonist's mental states (both implicit and explicit inferences, also imputed by readers) and psychological trajectory correlate with the classic dramatic arc of stories. Thus, to obtain the protagonist's mental states, we refer to a recent work [VR21b; Sap+19a; Ras+18c] that tracks characters' mental states using an external memory module and enables pragmatic enrichment of narrative sentences based on social commonsense knowledge aggregated using information retrieval and data mining strategies. Towards our overarching goal of detecting climax and resolution in short personal narratives, we implement an end-to-end computational model that uses a multi-feature fusion technique to effectively integrate the protagonist's mental state representation with linguistic information at syntactic and semantic levels. Our contributions are summarized below:

- A STORIES[1] corpus containing a collection of Reddit Personal Narratives with fine-grained annotations of prominent structural elements of a narrative – climax and resolution.

- An end-to-end neural network for modeling narrative structure, referred to as M-SENSE[2], that allows for integration of protagonist's mental state representations with linguistic information through a multi-feature fusion technique.

[1]Short for **ST**ructures **O**f **R**eddI**t** P**E**sonal **S**tories
[2]Short for **M**ental **S**tate **E**nriched **N**arrative **S**tructure mod**E**l

- Experiments that analyze the impact of our modeling choices for short personal narratives. Specifically, we gauge the influence of incorporating mental state embeddings and report an improvement in F_1 scores of $\sim$ 11% and $\sim$ 13% over the base model for predicting climax and resolution, respectively.

6.2 Related Work

There is a large body of prior work that focuses on different aspects of narrative comprehension. Computational analysis of narratives operates at the level of characters and plot events. Examples include plot-related studies – story plot generation, plot summarization, detecting complex plot units, modeling event schemas and narrative chains and movie question-answering; character-based studies – inferring character personas or archetypes, analyzing inter-personal relationships and emotional trajectories, identifying enemies, allies, heroes [CVR18; BOS13; VCR20; Ras+18c; VR21b]; story-level analysis – story representation, predicting story endings, modeling story suspense, and creative or artistic storytelling, to list a few [VVZO15; LDL19; SCM16; Joc13; Fin16; Tam+18].

Several studies have analyzed the literature in narratology and formulated different goals and annotation labels associated with narratives towards modeling their structure. Elson's [Els12a] Story Intention Graph (SIG) provided an annotation schema to capture timelines as well as beliefs, intentions, and plans of story characters. The annotations in this approach are similar to story generation methods described in Belief-Desire-Intention agents [RG+95], and intention-based story planning [RY10]. Previous studies like [GBS11] have analyzed the personal web blog stories containing the everyday situation. Rahimtoroghi et al. [Rah+14], and Swanson et al. [Swa+14]

used a subset of Labov's categories, including orientation, action, and evaluation in such personal weblog narratives. Black and Wilensky (1979) evaluate the functionality of story grammars in story understanding, As explained earlier, [PKL19]'s dataset for analyzing turning points is a valuable addition in this area of work. Moreover, there has been consistent efforts [Jor+18; Jor+19] that study the link between Information Retrieval (IR) and narrative representations from text. These include works that exploit narrative structure in movies for IR [Jha08], detect and retrieve narratives in health domain (patient communities & medical reports) [Joh+08; RMA04; DVK19; KCZ17], identify narrative structures in news stories [BBP20; Lev+20] or generate summaries from screenplays or novels [Pap+20], to name a few. Given this broad spectrum of work, we leverage mental state representation models that are pretrained using social commonsense knowledge aggregated using IR and text mining techniques. We employ the ensuing mental state embeddings in tandem with contextual semantic embeddings towards our primary objective of identifying elements of high-level narrative structure – climax and resolution. We also conduct a detailed analysis of the outcome and the contribution of the protagonist's psychological state trajectory for our problem at hand.

6.3 Dataset

As described in detail in Chapter 3, we construct a manually annotated corpus, referred to as our STORIES dataset. This dataset contains a total of 2,382 Reddit personal narratives, comprising 42,614 sentences. Table 6.1 shows the descriptive statistics of our dataset. With our annotation setup, we are able to obtain a substantial inter-annotator agreement for both the categories indicating the rise and fall of dramatic tension (climax and resolution, respectively). Notably, the annotators can discern between the two interest

Dataset Statistics	
#Stories	2,382
#Total Sents.	42,614
#Climax Sents.	5,173
#Resolution Sents.	4,502

TABLE 6.1: Statistics of our annotated STORIES dataset.

categories despite the high cognitive load and complexity involved in extracting them from unstructured user-generated content. We will be using this dataset for the study in this work.

6.4 M-SENSE: Modeling Narrative Structure

In this work, we explore different modeling and analysis methods for understanding narratives and automatically extracting text segments that act as key elements of narrative structure, particularly climax and resolution. We use the collected dataset to train and evaluate models to identify sentences in a narrative that qualify as climax and resolution. The models are provided a narrative text T with L sentences, $T = [S_1, S_2, ..., S_L]$, as input. Here, each sentence S_i contains N_i words $\{w_1^i, w_2^i, .., w_{N_i}^i\}$ from vocabulary $\mathcal{V}$. Towards automatic detection of structural elements, we formulate it as a sentence labeling task where the goal is to predict a label $\hat{y}_i \in \{None, Climax, Resolution\}$ for each sentence S_i, based on the story context. Various modeling techniques examined in this work involve processing narratives at the sentence level. Beyond linguistic features extracted from narratives, we focus on a dominant aspect in which a narrative is formed or presented, which accounts for characters' mental states – goals, intentions, actions, and emotions. Thus, we leverage transfer learning from pretrained models trained to infer characters' mental states (specifically, intents and emotional reaction) from a narrative. Combining the embeddings extracted from pretrained mental state

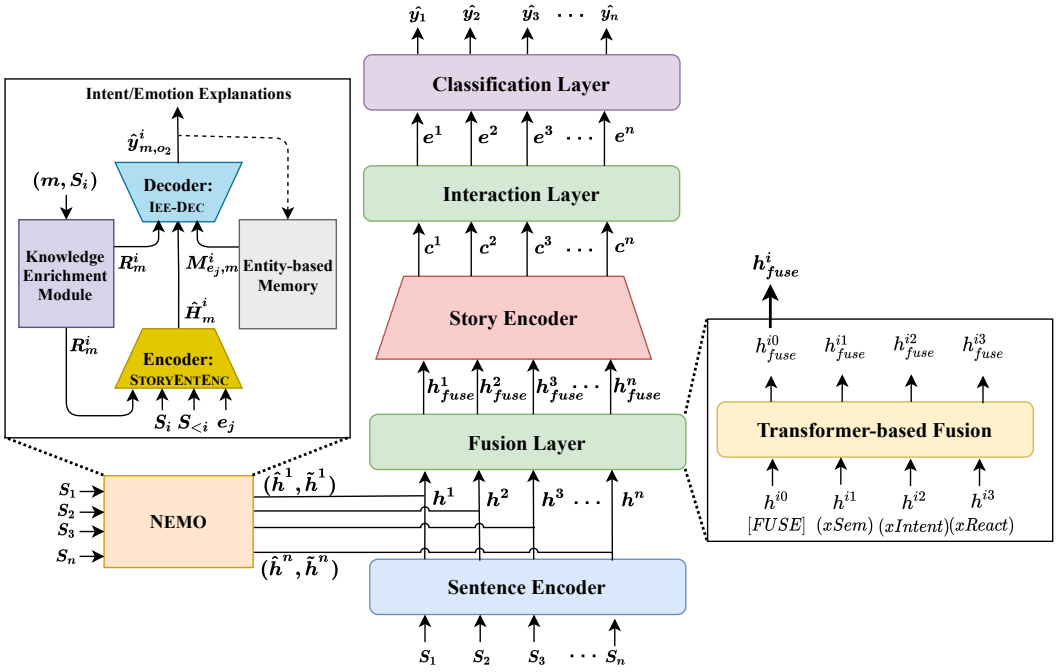


FIGURE 6.2: Illustration of our M-SENSE model. Note that $h^{i1} = h^i; h^{i2} = \hat{h}^i; h^{i3} = \tilde{h}^i$ relate to semantics $(xSem)$, intents $(xIntent)$ and emotional reactions $(xReact)$ of the i^{th} sentence respectively.

model with other semantic features may reap benefits from the relationship between characters' underlying psychological processes (or mental states) and narrative structure [Rya91; Pal10] and prior training involving social commonsense knowledge. Therefore, we investigate a multi-feature fusion-based learning model, M-SENSE, that potentially encapsulates syntactic, semantic, characters' mental state features towards our overall goal of predicting climax and resolution in short personal narratives. Below, we discuss the model components in more detail.

Our M-SENSE model comprises of the following components:

- **Ensemble Sentence Encoders**, which utilizes multiple encoders to produce per-sentence linguistic and mental state embeddings.

- **Fusion layer**, which integrates the protagonist's mental state information with the extracted linguistic features from sentences.

- **Story Encoder**, which maps the fused sentence encodings into a sequence of bidirectionally contextualized sentence embeddings. The input to this component can be either a sequence of token or sentence embeddings, depending on the modeling choice. Considering the importance of the overall story context towards evaluating each sentence's role in the entire narrative, this component integrates the surrounding story context information for computing rich sentence-level embeddings.
- **Interaction layer**, which estimates state transition across sequential context windows to identify the structure boundaries.
- **Classification layer**, which involves linear layers to calculate the label probabilities eventually.

6.4.1 Ensemble Sentence Encoders

Several sentence encoding methods [Dev+18; Cer+18; PSM14] have been proposed to tackle specific tasks or produce embeddings generalizable across multiple NLP problems. In this work, we aim to exploit both linguistic and cognitive or mental state features towards building an enhanced model for narratives. The former is extracted using a general-purpose language representation model usually trained on extensive text data (e.g., BERT, USE). In contrast, we use a dedicated pretrained task-specific model for the latter.

Extracting Linguistic Representations

Pretrained general-purpose sentence encoders usually capture a hierarchy of linguistic information such as low-level surface features, syntactic features, and high-level semantic features. Given a narrative text with L sentences $T = [S_1, S_2, ..., S_L]$, this component outputs hidden representations for sentences $H_{sents} = [h^1, h^2, ..., h^L]$ using different encoding methods. We intuit that the

choice of sentence encoding models and features extracted can significantly impact the overall performance for the task at hand. Instead of training a sentence encoder from scratch, we leverage pretrained models to produce embeddings for sentences in the narrative. In our MSENSE model, we use a token-level BERT-based sentence encoder.

Token-level BERT: Since BERT produces output vectors that are grounded to tokens instead of sentences, we undertake an *input processing step* that involves insertion of a special $[CLS]$ token at the beginning of each sentence and a $[SEP]$ token at the end of each sentence in the input sequence. we feed the entire narrative text T to the input processing step to get the following sequence: $\{[CLS], w_1^1, w_2^1, .., w_{N_1}^1, [SEP], [CLS], w_1^2, w_2^2, .., w_{N_2}^2, [SEP], .., [SEP], [CLS], w_1^L, w_2^L, .., w_{N_L}^L, [SEP]\}$, where w_j^i is the j^{th} in i^{th} sentence in the narrative. The multiple $[CLS]$ symbols will aggregate the features for sentences taking the context into consideration. Next, we apply alternating segment embeddings indicative of different sentences in our input textual narrative. Given a narrative with four sentences, $[S_1, S_2, S_3, S_4]$, we assign the segment embeddings as $[E_A, E_B, E_A, E_B]$. Finally, this processed input is fed to the pretrained BERT model as:

$$H = [h_{[CLS]}^1, .., h_{N_1}^1, h_{[SEP]}^1, .., h_{[CLS]}^i, .., h_{N_i}^i, .., h_{[SEP]}^L] = \text{BERT}(T) \tag{6.1}$$

The hidden representation of i^{th} $[CLS]$ token from the top BERT layer is extracted as the semantic embedding of the i^{th} sentence. However, we drop the subscript $[CLS]$ from $h_{[CLS]}^i$ and denote the output semantic embeddings as:

$$H_{sents}^{xSem} = [h^1, h^2, ..., h^L] \tag{6.2}$$

In section 6.7, we compare the performance of token-level BERT with other sentence encoding methods, including sentence-level BERT and USE [Cer+18], that process each sentence independently.

Incorporating Protagonist's Mental Representation

Prior studies have explored how a story's progression is as much a reflection of a sequence of a protagonist's motivation and emotional states as it is the workings of an abstract grammar [Pal02; Moh13; AS05]. Thus, it is reasonable to assess the role of cognitive tension that the characters grapple with beyond linguistic patterns and compute the change in the protagonist's psychological states for automatically determining the structural components of the narrative.

Following a recent work [VR21b; Ras+18c] that implements a NEMO[3] model, a variant of a Transformer-based encoder-decoder architecture, to embed and explain characters' (or entities) mental states. We extract the embeddings of intents and emotional reactions of the protagonist for a given sentence in the narrative conditioning on the prior story context. Figure 6.2 contains the overview of NEMO architecture also. The computation of mental state embeddings are facilitated by a knowledge enrichment module that consolidates commonsense knowledge about social interactions and an external memory module that tracks entities' mental states. The social commonsense knowledge is aggregated using information retrieval and extraction techniques. Using prior context $(S_{<i})$, entity (e_j) and mental state attribute information ($m \in \{xIntent, xReact\}$ representing intent and emotional reaction respectively), we use the encoder, STORYENTENC$(\cdot)$, in this trained model to obtain entity-aware mental state representation of the current sentence S_i. The encoding process in the NEMO model is given by:

$$(\hat{H}^i_{xIntent}, \tilde{H}^i_{xReact}) = \text{STORYENTENC}(S_i, S_{<i}, e_j, m); \quad \forall m \in \{xIntent, xReact\} \quad (6.3)$$

[3]Narrative Entity Mental Model

where $e_j \in \mathcal{E}$ is the entity, $(\hat{H}^i_{xIntent}, \tilde{H}^i_{xReact})$ is the resulting entity-aware intent and emotion representation of the i^{th}-sentence given the story context. In this work, we use the narrator ("I" or "self" in the personal narratives) as the protagonist. We only utilize the hidden representations of the $[CLS]$ token from both $(\hat{H}^i_{xIntent}, \tilde{H}^i_{xReact})$ for subsequent processing steps. We denote these intent and emotion representation as:

$$H^{xintent}_{sents} = [\hat{h}^1, .., \hat{h}^L] \tag{6.4}$$

$$H^{xReact}_{sents} = [\tilde{h}^1, .., \tilde{h}^L] \tag{6.5}$$

6.4.2 Transformer-based Fusion Layer

Given sentence representations computed using multiple methods, we apply a fusion strategy that weighs the relevance of each latent vector and derives a unified sentence embedding for our classification task. Let $h^{ik}; \forall k \in \{1, ..., K\}$ denote different per-sentence latent vectors. In our case, $K = 3$ and $h^{i1} = h^i; h^{i2} = \hat{h}^i; h^{i3} = \tilde{h}^i$ are embeddings related to semantics $(xSem)$, intents $(xIntent)$ and emotional reactions $(xReact)$ of the i^{th} sentence respectively.

Drawing ideas from the literature of multimodal analysis [UMS+20], we treat the multiple latent vectors as a sequence of features by first concatenating them together. We introduce a special token $[FUSE]$ [4] that accumulates the latent features from different sentence encodings. The final hidden representation of $[FUSE]$ token obtained after feeding them to a Transformer layer is the fused output sentence representation.

$$h^i_{fuse} = \text{TF}(\|^K_{k=0} h^{ik}) \tag{6.6}$$

where TF refers to the transformer encoder layer and h^{i0} (i.e. when $k = 0$) is set to the trainable $[FUSE]$ vector.

[4] $[FUSE]$ is similar to the commonly used $[CLS]$ token.

6.4.3 Story Encoder

Understanding key elements of the narrative structure requires aggregation of different narrative-level features such as important events and characters in the story, their relationships, the impact they have on characters' mental states, and their transitions, to name a few. Since we use pretrained models for embedding the sentences in the narrative, we introduce story encoding layers dedicated to capture these narrative-level features by combining relevance-weighted inter-sentence features bidirectionally. Given the sentence embeddings obtained from the sentence encoder, $H_{sents} = [h^1, h^2, ..., h^L]$, we compute contextualized sentence embedding $c^i; \forall i \in \{1, 2, .., L\}$ by conditioning on the surrounding context sentences in the narrative. These contextualized embeddings will be eventually used for predicting the sentence labels (Climax, Resolution, or None).

We apply Transformer layers on the top of the sentence representations to extract narrative-level features focusing on the task of detecting elements of high-level narrative structure. We refer it as *Inter-sentence Transformer*. Intuitively, Transformer layer involves the multi-head attention mechanism that: (a) focuses on possibly different sentences in the narrative, and (b) produces context-aware sentence embedding by combining the features across these sentences to decide if the corresponding label should be assigned to the sentence under consideration. This is given as:

$$\begin{aligned} \hat{H}^l &= LayerNorm(\hat{C}^{l-1} + \textsc{Mha}(\hat{C}^{l-1}) \\ \hat{C}^l &= LayerNorm(\hat{H}^l + \textsc{Ffl}(\hat{H}^l)) \\ C_{sents} &= [c^1, c^2, ..., c^L] = \hat{C}^{n_L} \end{aligned} \tag{6.7}$$

where $\hat{C}^0 = PE(H_{sents})$, PE refers to the positional encoding applied to inject order information to the sentence vectors, H_{sents} contains the sequence of sentence vectors output by our sentence encoder module, $LayerNorm$ refers

to layer normalization operation, MHA is the multi-head attention operation and FFL is a feed-forward layer [Vas+17]. The superscript l indicates the depth of the stacked Transformer layers. The output from the topmost layer, $l = n_L$, is our contextualized sentence embeddings C_{sents}. In section 6.7.2, we compare the relative advantage of using a Transformer-based story encoder over an RNN-based encoding method.

6.4.4 Interaction layer

Inspired by traditional segmentation approaches [Hea97], prior works have utilized interaction layer to determine topic boundaries or thematic units in movie synopsis [PKL19], question-answering, and document matching tasks, to list a few. Consistent with these studies, we reckon that the local interaction information with the surrounding narrative context can be useful to determine the boundaries of various elements of the narrative structure. In this layer, we compute the transition of state across sentences by measuring similarity in the embedding space between sequential context windows. By choosing windows of size s, we compute the left (c^i_{left}) and right (c^i_{right}) context information for the i^{th} sentence by computing the mean sentence embedding within that window. Features representing different similarity measures such as element-wise product, cosine similarity, and pairwise distance are computed for both left and right mean context representation. The similarity metrics are concatenated along with contextualized embeddings (sentence, left & right context) to get interaction-feature enhanced context-aware embeddings:

$$E_{sents} = [e^1, e^2, ..., e^L] \tag{6.8}$$

6.4.5 Classification layer

We employ linear layers, f_s, on the top of interaction layer outputs and maps the enriched sentence embedding to a C-dimensional output. Here, $C = 3$ is the number of classification labels. We apply a softmax activation to get the probability distribution over the sentence labels/categories associated with narrative structure. This step is given as:

$$\hat{y}_i = softmax(f_s(e^i)) \tag{6.9}$$

where f_s involves linear layers that map the enriched sentence embedding to a k-dimensional output in addition to straightforward methods like element-wise multiplication, element-wise summation, and concatenation with a linear layer for our task.

6.5 Zero-shot Approaches

In addition to our M-SENSE model and its variants, we experiment with zero-shot methods that utilize either simple heuristics or suspense-based approaches to model narrative structure.

Heuristic-based Approaches

In addition to different modeling approaches, we experiment with simple heuristics for automatically labeling the sentences in the story. This method assumes that the title of the Reddit post provides the summary of the post and hence, could refer to the MRE/climax of the narrative. We use a pre-trained sentence embedding model and compute the semantic similarity between each sentence in the narrative and the original title of the Reddit post. We calculate USE embeddings of sentences to identify the nearest neighbor of the post title and label it as the climax. Next, we assign the last sentence

as the resolution because it is more like for the cognitive tension to drop and reach a resolution at the near end of the narrative.

Suspense-based Approaches

A recent work [WK20] has explored surprise and uncertainty reduction as a measure of suspense in narratives by considering sentences as the primary unit of processing. In our work, we mainly focus on surprise values based on consequential state change in narratives. Intuitively, a large difference in any particular state indicates increased surprise at that point in the narrative. [EFK15]'s surprise is defined as the amount of change from the previous sentence to the current sentence in the narrative. Peaks in such measures could reflect potential events where the protagonist faces principal obstacles, and these may act as the defining moments of narrative structure. Thus, we examine different sentence embedding techniques to compute state changes, including change in the protagonist's mental state representations in our experiments to recognize suspenseful states. This will act as a relevant baseline to determine the effectiveness of semantic and mental state features.

6.6 Training & Hyperparameters

We approach the narrative structure model as a sentence classification task. We divide the collected data based on the number of narratives into the train, validation, and test sets at 70-10-20 split. In our sentence labeling task, each sentence will be accompanied by the entire narrative context. On the training set, we perform data augmentation by replacing sentences in narrative context with their paraphrases. The paraphrases are generated using a back-translation approach [Edu+18] based on pretrained English↔German translation model. We limit the number of such modified sentences to 20% of the story length. We tune the hyperparameters using grid search, and the

best configuration is obtained based on validation set performance. Our best configuration consists of a two-layer story encoder ($n_L = 2$) and a single transformer-based fusion layer with 12 heads. We also added a dropout with a rate of 0.2 to prevent overfitting. We optimize using Adam [KB14] at a learning rate of $\alpha = 0.0001$ and a batch size of 32 with PyTorch [Pas+19] being our model implementation framework.

6.7 Experiments

We conduct experiments to study the following research questions:

RQ1: How does our model compare with other baselines for identifying climax and resolution in short personal narratives?

RQ2: How do various model components contribute to the overall performance? To what extent do mental state representations play a role in our classification task?

6.7.1 Overall Predictive Performance (RQ1)

Baselines

We conduct experiments that evaluate our model compared to different prior approaches [PKL19; WK20]. We compare our model with a set of carefully selected zero-shot (Section 6.5) & supervised baselines, shown as follows.

- **Random baseline**, which assigns labels (Climax, Resolution or None) to sentences randomly.
- **Distribution baseline**, which picks sentences that lie on the peaks of the empirical distributions for climax and resolution in our training set as explained in Section 3.2.2.

- **Heuristic baseline**, which labels the sentences as climax or resolution based on heuristics as in section 6.5. While we use the sentence that is the closest semantic neighbor of the post title as the climax, the last sentence in the narrative is labeled as the resolution.

- **GloVeSim** [PSM14], which measures the cosine similarity between the mean GloVe word embeddings of two sentences. This estimates the change features and acts as an alternative to Ely's surprise measures.

- **StoryEnc** [PKL19], which uses the hierarchical RNN based language model to encode sentences in the story and eventually compute suspense measures for our classification task.

- **Bert** [Dev+18], which embeds the narrative sentences as in section 6.4.1 and then suspense-based approach is used.

- **Use** [Cer+18], which encodes sentences and utilizes the resultant embeddings to compute suspense measures for our task at hand.

- **StoryEntEnc** [VR21b], which encodes the sentences in the story from the protagonist's perspective. Here, we denote intent and emotional reaction embeddings as $(E_{int} = H_{sents}^{xIntent})$ and $(E_{emo} = H_{sents}^{xReact})$ respectively. We utilize these embeddings for measuring suspense and applying them for our classification task.

- **Cam** [PKL19], which consists of bidirectional LSTM model stacked on the top of the sentence embeddings to obtain contextualized representations. Cam is the abbreviation for context-aware model.

- **Tam** [PKL19], which uses a RNN-based contextualized sentence encoder enriched with interaction layer to compute boundaries between different topics or thematic units in stories. This model is referred to as the topic-aware model (Tam).

Models	$F_1 \uparrow$		$D \downarrow$	
	C	**R**	**C**	**R**
Random baseline	0.196	0.143	29.05	30.57
Distribution baseline	0.274	0.315	**15.79**	**14.42**
Heuristic baseline	0.217	0.147	23.74	26.82
GloVeSim	0.312	0.344	12.06	11.65
Token-level BERT	0.408	0.441	9.37	8.09
Sentence-level BERT	0.352	0.366	10.88	9.73
Sentence-level USE	0.379	0.391	10.42	9.58
STORYENC	0.410	0.438	8.81	7.46
E_{int}	**0.437**	0.462	**8.19**	6.94
E_{emo}	0.429	**0.475**	8.43	**6.67**
TAM	0.565	0.609	5.90	5.02
CAM	0.578	0.604	6.58	5.44
M-SENSE	**0.694**	**0.743**	**4.15**	**3.20**

TABLE 6.2: Evaluation Results of different models for detecting climax (**C**) and resolution (**R**) in short personal narratives. We report F_1 score per class & percent mean annotation distance (D) for these models. We use $\uparrow, \downarrow$ to indicate if higher or lower values mean better performance respectively.

- **M-SENSE–FUSION**, which is a variant of our M-SENSE model without mental state embeddings. This means that we remove the components – NEMO and Fusion layer in Figure 6.2.

- **M-SENSE**, which is our complete model incorporating protagonist's mental representation as described in section 6.4.

Results

Table 6.2 outlines the results of our evaluation. We report the performance of simple baselines, of which the distribution baseline turns out to be the strongest. Though heuristic baseline performs poorly, we find a marginal performance improvement compared to a random baseline. This is reflected more in the distance measure (%) D than the F_1 score. This suggests that the Reddit post title does contain some relevant signal to better predict the climax

in the narrative. At the same time, the last sentence heuristic for resolution is only as good as a random classifier.

Applying suspense-based approaches with different sentence embedding methods yields relative improvement over the simple baselines in terms of both the evaluation metrics. As expected, sentence-level BERT/USE performs poorly than its token-level counterpart. We attribute this variation in performance to the lack of any story context information for computing latent embedding, thereby affecting the assessment of state changes in the narrative. However, sentence-level USE's ability to produce better similarity estimates gives it a slight advantage over sentence-level BERT. Notably, sentence representations obtained from models trained on stories (STORYENC, STORYENTENC) recorded comparable to improved results over other sentence embedding methods. Strikingly, computing surprise using protagonist mental state embeddings exhibits an overall enhanced classification capability. We find that the intent embedding (E_{int}) helps achieve the best zero-shot performance for detecting climax. Competitive outcome for resolution is obtained using protagonist's emotion representation (E_{emo}).

We compare our complete M-SENSE model with the best performing prior models such as CAM, TAM [PKL19] applied for similar tasks. As we can see, supervised fine-tuning approaches easily beat the earlier results obtained using zero-shot methods. Though CAM has better F_1 score for climax prediction, TAM outperforms CAM in terms of D for predicting both the climax and resolution. Finally, our M-SENSE model achieves an absolute improvement of $\sim$ 20.07% and $\sim$ 22% for climax and resolution prediction respectively. The factors that aid us towards actualizing this performance include – (a) integration of protagonist's mental state features via fusion layers, and (b) long-term story contextualization using modeling choices like Transformer-based story encoder.

Model Variants	$F_1 \uparrow$	
	C	**R**
M-SENSE	**0.688**	**0.738**
Sentence Encoder Variants		
w/ Sentence-level BERT	0.665	0.709
w/ Sentence-level USE	0.677	0.726
Story Encoder Variant		
w/o Story Encoder	0.620	0.653
w/ Inter-Sentence RNN	0.659	0.705
Interaction Layer Variant		
w/o Interaction Layer	0.654	0.716
Fusion Layer Variants		
–w/o Fusion Layer	0.614	0.640
–w/o E_{int}	0.638	0.703
–w/o E_{emo}	0.652	0.687

TABLE 6.3: Ablation Results: We report F_1 score per class (**C**limax and **R**esolution) with non-default modeling choices for individual components of our M-SENSE model.

6.7.2 Ablation Study (RQ2)

To evaluate each component's contributions in our M-SENSE model, we conduct an ablation study using the validation set. For this study, we compare our best performing M-SENSE model with alternative modeling choices for each of the components. Table 6.3 shows the results of our study. We modify one component at a time and report their corresponding performance using F_1 metric. This involves either replacing a component (denoted by "w/") or removing a component (denoted by "w/o" to refer without the component). For eg. "w/ Sentence-level BERT" refers to replacing token-level BERT in default M-SENSE model with sentence-level BERT as our sentence encoder; "w/o E_{emo}" indicates the removal of protagonist's emotion state embedding from the fusion layer.

Choice of Sentence Encoder: From the results, it is clear that token-level BERT generally performs better the sentence-level BERT variant. This is unsurprising as the sentence-level approach produces embeddings without story

context. Such an approach results in the loss of fine-grained inter-sentence token dependencies that the token-level BERT can extract. Experiments also suggest that sentence-level USE model trained on textual similarity tasks results in better F_1 scores than the BERT counterpart.

Importance of Contextualization: Next, we evaluate the contribution of a story encoder to our classification task. We observe that the performance drops significantly (by $\sim$ 10% and $\sim$ 12% for the climax and resolution prediction, respectively) without a story encoder. The importance of contextualizing the story sentences is established as we see a marked improvement of $\sim$ 8% in the average overall performance with the introduction of an inter-sentence RNN story encoding layer. Still, this performance lags behind our default M-SENSE setting with the Transformer-based story encoding layer. We find the story encoder relevant even if the inter-sentence dependencies are captured using the token-level BERT model. We attribute this to the task-specific inter-sentence relationships being unearthed as we fine-tune our model.

Impact of Interaction Layer: The addition of an interaction layer yields an average $\sim$ 4% gain in performance for identifying climax and resolution. The advantage of introducing an interaction layer has been studied in prior studies [Hea97; PKL19] and we find the performance improvement to be congruous with these studies.

Influence of Mental State Embeddings: The purpose of a fusion layer is to incorporate sentence level embeddings capturing the protagonist's intent and emotional reactions. In this study, we examine the necessity of a fusion layer and probe the influence of the protagonist's mental state embeddings on our classification task. Notably, the results in Table 6.3 validate the benefits of introducing the fusion layer and demonstrate the relative performance gains obtained with intent and emotion embeddings. In the absence of a fusion

layer, we observe that the performance drop is $\sim 11\%$ and $\sim 13\%$ for predicting climax and resolution, respectively. The loss of the protagonist's intent information impacts the climax prediction more. This is analogous to the effect emotion information has on resolution prediction. In Section 6.7.3, we delve deeper to analyze their relative importance.

6.7.3 Analysis and Discussion

Effect of Story Length: We conduct experiments to analyze how different sentence encoders and mental representation fusion impact the overall performance. For this analysis, we compare different sentence encoders' performance with and without fusion layer for detecting climax in narratives with a varying number of sentences (Story length). Figure 6.3 shows the results of this analysis. We observe that the token-level BERT outperforms sentence-level BERT and USE encoders for narratives containing up to $13 - 14$ sentences, but the performance gradually degrades beyond 14 sentences. Sentence-level USE encoder produces a stable and relatively better outcomes for longer narratives (story length > 14). With the introduction of mental state representation through the fusion layer, the F_1 score improved significantly irrespective of the sentence encoder used. However, we find that the token-level BERT and USE enriched with mental state embeddings yielded a comparable performance, with the former having a slight edge over the latter. Also, the performance degradation of token-level BERT is mitigated as the fusion layer is added, and this is reflected in the F_1 score even for longer narratives. Thus, this analysis revalidates the use of our modeling choices in our M-SENSE model.

Error Analysis: In order to estimate why our model augmented with mental state representation performs better, we conduct error analysis between

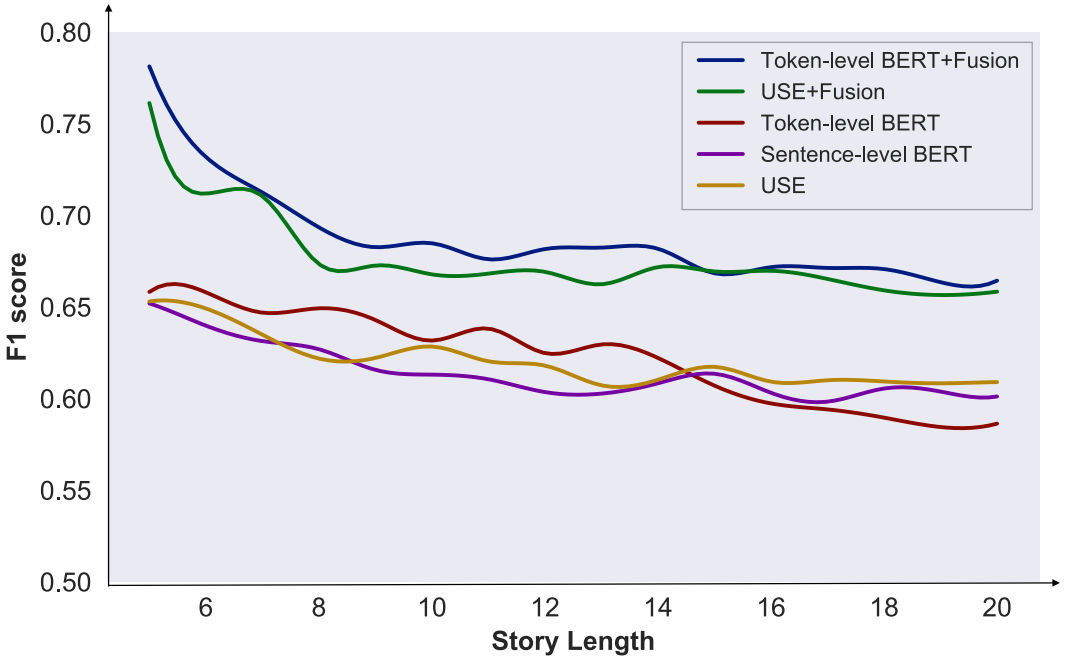


FIGURE 6.3: Performance of sentence encoders for detecting climax in story with varying length.

our full M-SENSE model and the model without mental representation fusion (M-SENSE $-$ *Fusion*). For those narratives where the latter model fails to predict correctly, we gauge the patterns emerging out of the following analysis: (a) Using VADER[5] [HG14] a normalized, weighted composite sentiment score is computed for each sentence in the narrative. Use typical threshold values used in the literature, we categorize these sentences as positive, neutral, or negative, and eliminate stories containing neutral sentences in the neighborhood of ground truth sentences, and (b) Using state classification [Ras+18c; VR21b], we assess Maslow's motivation or intent categories associated with sentences predicted as climax or resolution in the narrative and analyze for any pattern related to ground truth climax/resolution sentences. For predicting resolution, the M-SENSE $-$ *Fusion* makes 28% more mistakes than M-SENSE model for narratives with homogeneous endings (i.e., narratives having the same sentiment sentences in the neighborhood of resolution

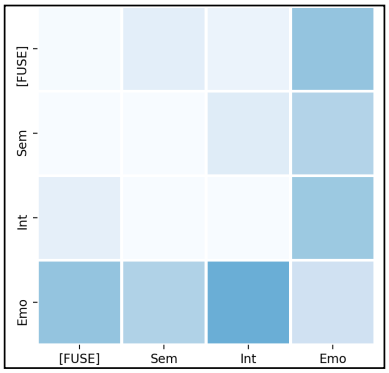
[5]https://github.com/cjhutto/vaderSentiment

closer to the end of the story). M-SENSE − *Fusion* model is unable to discern clearly and predicts a different sentence as resolution. Based on our analysis (b), there is a clear pattern that M-SENSE gains significantly over the M-SENSE − *Fusion* when the ground-truth climax sentences belong to "Esteem" and "Love/Belonging" categories.

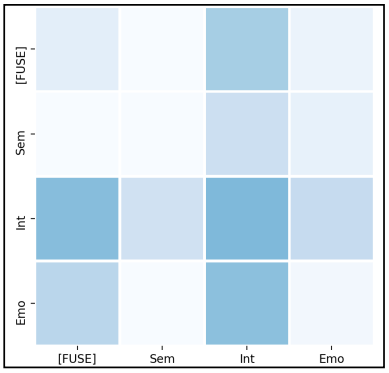
Attention Analysis: We conduct attention analysis on those narratives containing sentences belonging to "Esteem" and "Love/ Belonging" categories. Specifically, we study the functioning of the Transformer-based Fusion layer for aggregating multiple latent embeddings – Semantic (Sem), Intent (Int), and Emotion (Emo). We then visualize the attention heatmaps from the fusion layer corresponding to these predicted climax and resolution sentences by computing the average attention score map over all heads. Figure 6.4 displays the visualized attention map for sample stories belonging to the above-mentioned categories. Since $[FUSE]$ is the aggregated output over all the three latent embeddings. We note that the attention map has high attention scores between intent (Int) and $[FUSE]$ vectors for stories related to "Esteem" motivation category, while more weight is assigned for emotion (Emo) in samples associated with "Love/Belonging" category.

6.8 Task: Modeling Movie Turning Points

Given that our work is primarily focused on modeling narrative structure in personal narratives, we analyze how we can apply such a model towards identifying climax and resolution in movie plot synopsis. Recent work by [PKL19] introduced a TRIPOD dataset containing a corpus of movie synopses annotated with turning points (TPs). By testing our model on this dataset, we evaluate our model's performance on an out-of-domain dataset. The dataset identified five major turning points in the movie synopses and screenplay, referring to them as critical events that prevent the narrative from drifting



I've been a lonely person all through my life due to my dysfunctional family. Though I've a brother who lives close to my place, I've always assumed we are totally different people. We don't connect very well at all. My life was a mess and always craved for some emotional and moral support more than financial. **I felt dejected and not worth living. So I attempted suicide using sleeping pills.** But now I look back at it and I think suicide attempt did change life for the better. **My brother saved me that day and he is my best friend that I love him to bits now. We realized our mistakes.** And how both of us were going through similar issues and being together was the best thing we could do to each other. I just wish I had bonded with him earlier.



I'm a manager in the retail store I work in, but it's more of an assistant manager position. I've been working at the same place for about a year. I actually left my job for 3 months to try a different job and make some more money. That job was terrible, and my current job welcomed me back with open arms. There are 3 main managers. I've always looked up to them. **Since one of the managers was moving out in a few months, there was an opening and I got completely motivated by the role. I've been training hard for the role and achieve my long term target.** My friends never believed in my capabilities. I wanted to break my friends' notion about my abilities. **And today, I've hit the bull's eye and excited that I'will take over as the manager for the new role.** I'm not sure how my friends will feel but I'm not going to brag about it. I just bought a new house, I'm happy it will make things a lot easier. I'm really excited for the new job.

FIGURE 6.4: Attention analysis of two stories with climax and resolution sentences related to Maslow's categories – Esteem (top) and Love/Belonging (bottom).

away. These five TPs are: Opportunity (TP1), Change of Plans (TP2), Point of No Return (TP3), Major Setback (TP4), and Climax (TP5). By their definitions for each of these categories [PKL19], TP4 and TP5 align clearly with our usage of climax and resolution from prior narrative theories. Due to this alignment, it is relevant to use our model to predict these two categories in the TRIPOD dataset. However, we focus on the movie plot synopses in this work and use the cast information collected from IMDb as a part of this dataset.

We first apply our MSENSE trained on our STORIES corpus directly and evaluate its zero-shot performance. We refer to this as a zero-shot approach given by (ZS) We assume the protagonist in the movie to be the top character from the IMDb cast information. Though this may not always be true, it measures how our model fares on this dataset for predicting TP4 and TP5. Further, we use sentence-level USE-based sentence encoder as some of the wiki

plot synopses are longer than what can be accommodated by our token-level BERT model. Additionally, we also fine-tune our model with the training set of the TRIPOD dataset. This is denoted by MSENSE(FT).

6.8.1 Results

We display our model's performance compared to the best performing TAM reported in the original work [PKL19]. The topic aware model TAM with TP views implemented separate encoders for each of the categories and computed different representations for the same sentences acting as different views related t each TP. Similarly, TAM+Entities enriched the model with entity information by applying co-reference resolution and obtain entity-specific representations. We compare these models without MSENSE model in zero-shot and supervised settings. Table 6.4 shows our model's results compared to the prior proposed approaches for modeling turning points in plot synopses. We find that our model in zero-shot settings outperforms a supervised TAM+TP views model, though it falls slightly behind the best supervised model in terms of the mean annotation distance (%). We intuit the advantage of training on short personal narratives to be providing us the edge on this model. It could be a bigger challenge if we have to identify other turning points in the story. Also, we restrict our model for predicting only two of the five major turning point labels. Finally, our fine-tuned model outperforms the best performing model, significantly reducing the mean annotation by an average of $\sim$ 20% on both the turning point labels. Thus, we are able to achieve remarkable improvement on an out-of-domain dataset even with assumptions on protagonist information. Therefore, we demonstrate that our MSENSE model can predict climax and resolution in stories beyond just personal narratives, albeit limited by story length at this point. We believe that there is tremendous scope for drawing insights by analyzing

Methods	**TP4**	**TP5**
TAM+TP views	6.91	4.26
TAM+Entities	5.23	3.48
MSENSE(ZS)	6.62	4.54
MSENSE(FT)	**4.17**	**2.38**

TABLE 6.4: Results of evaluation on TRIPOD dataset. We report Mean Annotation Distance (%) D results for identifying TP4 and TP5 relevant to this work.

narratives across varied genres and performing interpretability assessments on the characters' mental states. However, we leave such an intricate piece of comparative analyses as future work.

6.9 Conclusion

Modeling high-level narrative structure through our modeling framework can facilitate future research towards nuanced discourse analysis and formal representation necessary for narrative retrieval, automated reasoning, or narrative generation. Towards this goal, we construct a dataset of personal narratives from Reddit containing annotations of climax and resolution sentences. Using our annotation setup, we are able to achieve a substantial inter-annotator agreement for both categories indicating the rise and fall of dramatic tension. Next, we address the challenge of automatically identifying these elements of narrative structure as a sentence labeling task. Understanding and quantifying the role of shifts in protagonists' psychological states and their interplay with semantic features are central to our research. Therefore, we introduce an end-to-end deep neural model, referred to as M-SENSE, that learns to effectively integrate the protagonist's psychological state features with linguistic information towards improved modeling of narrative structure. We experimentally confirm that our model outperforms several zero-shot and supervised baselines and benefits significantly from incorporating

the protagonist's mental state representations. Our model is able to achieve $\sim$ 20% higher success in detecting climax and resolution in short personal narratives than the previous methods. We believe that our work will advance the research in understanding the larger dynamics of narrative communication and aid future efforts towards developing interesting AI tools that can interact with human users through stories.

Chapter 7

Conclusion and Future Work

Stories have long been theorized to influence the human understanding of the social world and promote social cognitive capabilities. Recently, there has been a growing interest to conduct empirical research exploring the interplay between narratives and mentalizing using a wide range of approaches. This dissertation aims to investigate computational methods to leverage this interdependence towards a grand challenge of endowing human-level social skills to machines. By exploiting personal narratives produced by ordinary people on social media, which act as a great source of commonsense knowledge about social situations, we develop techniques that examine this mutual influence by (a) learning aspects of mentalizing (motives and emotions) through pragmatics-enriched social event embeddings and dynamic entity state tracking modules; and (b) demonstrating improved narrative comprehension through models that effectively predict prominent elements of narrative structure – climax and resolution using the mental state representations.

In the following sections, we delineate the key highlights of our contributions, along with the limitations of our work. Further, we lay down directions for future work that could potentially address those limitations and assess research areas where the models proposed in our work can offer a nudge towards enhanced performance.

7.1 Contributions

The key contributions of the work include identification of key functional modules that focus on leverage the relationship between mentalizing and narrative comprehension to capture specific aspects of aspects of social competence. Since we divide our work into three primary modular components, we explain the contributions of each of these components and how they eventually add up to our overarching goal.

7.1.1 Learning Knowledge-Enriched Social Event Representation

Events are considered the focal points of situations conveyed in narratives and connected in memory along various dimensions like time, space, characters (or protagonist), causality, and intentionality. Specifically, the events related to social situations demand an understanding beyond mere linguistic structures. Therefore, we make efforts to incorporate pragmatic properties referring to the knowledge outside the explicitly stated content in the event text. This knowledge is related to the human's inferred implicit understanding of event actors' intents, beliefs, and feelings or reactions. For this purpose, we aggregated sentence-level implicit intent and emotion states using web-search-based data mining techniques. We identified the potential issues with prior knowledge sources that replaced social role information with 'Person' markers leading to a loss of information. Therefore, we alleviate this issue with our aggregated social commonsense knowledge assertions (SB-SCK) that contain social role information in addition to the apparent intent and emotional reaction expressions related to the event text. Though the dataset is noisy, it provided ample opportunity to improve the learning of our social event embeddings. Using multiple knowledge sources like CONCEPTNET, ATOMIC and our very own SB-SCK, we sharpened the

social event embeddings to move beyond mere linguistic structures and embed information at syntactic, semantic, and pragmatic levels through the additional knowledge we feed into the model. We implemented a fine-tuned BERT-based model to disentangle pragmatic and non-pragmatic properties and eventually encapsulate them into a unified social event representation. By evaluating our work on a held-out test corpus of social events, we established the advantages of using such a model over several other baselines. Further, our experimental results showed improved performance on several downstream tasks like event similarity (Accuracy: 71.23%), reasoning (Accuracy: 67.9%), and paraphrase detection (Accuracy: 90.16%) tasks. These outcomes provided the necessary empirical evidence for their applicability in subsequent modeling challenges involving narratives and mentalizing in this work.

7.1.2 Modeling Human Motives and Emotions from Personal Narratives

Towards the goal of nurturing aspects of mentalizing abilities, we addressed a subset of challenges related to the lack of annotated data that allows for convincingly embedding and explaining characters' mental states. Further, we also investigated methods that can dynamically track mental states of characters throughout the story. We constructed a weakly annotated corpus of personal stories from Reddit using data mining and information extraction strategies to obtain expressions of motivation and emotional reactions. We implemented a NEMO model, which is a variant of Transformer-based architecture augmented with additional memory modules that can capture sentence-level event pragmatics from our EVENTBERT and enable dynamic entities' mental state tracking.

We demonstrated our proposed method's superior performance over several baselines in mental state classification and generation tasks. Our complete model achieved an absolute mean improvement of $\sim$ 8% and $\sim$ 11% over a fine-tuned GPT-2 model using the embedding average metric of the generated intent and emotion explanations respectively across two different story datasets. With the help of pretraining, we observed that the low-resource domains could benefit from the pretrianing of our model and transfer the knowledge to stories in other disciplines. Even in mental state classification tasks, our model exhibited significant leads over prior approaches with an F_1 score of $\sim$ 0.72 and $\sim$ 0.71 for Maslow and Plutchik categories, respectively. In order to evaluate the learned mental state embeddings, we apply these embeddings on a downstream empathetic response generation task model. Our mental state-enriched ensemble model with a relatively lower number of parameters resulted in significant improvement in automated metrics in the response generation task. Thus, we successfully show how narratives can be tapped in to construct models that showcase mentalizing capability, albeit in a limited capacity of predicting motives and emotional reactions only.

7.1.3 Modeling Narrative Structure in Short Personal Narratives

Given the substantial outcomes obtained using our mental state representation and generation framework, we determine the applicability of these mental representations for modeling high-level narrative structure in narratives. This requires the processing of the entire narrative to identify the boundaries of these structural components of the narrative. Through this work, we are able to identify the climax and resolution in the stories though there is ample scope for improvement. The vital contributions of this work lie both in the

data and modeling aspects of the work. We created a resource for future researchers containing reliable annotations of climax and resolution sentences in personal narratives obtained from Reddit from the data perspective. Using our annotation setup, we accomplished substantial inter-annotator agreement with Cohen's $\kappa = 0.651$ and $\kappa = 0.769$ for climax and resolution, respectively. The agreement/reliability scores indicate that our annotation setup allows for a clear separation of major structural elements of the narrative, capturing the dramatic rise and fall of the characters' tension in the story.

Regarding the modeling aspect of our work, we tackle the challenges involved by incorporating the protagonists' psychological states and situating them in the bidirectional story context. Using our M-SENSE model, the protagonist's psychological state features are integrated with linguistic information relying on embeddings obtained from pretrained language models. By tracking the shifts in these mental states and applying multi-feature fusion with linguistic features, we are able to achieve a per-class F_1 score of 0.694 and 0.743 for climax and resolution, respectively in the sentence labeling task. Since quantifying the role of shifts in protagonists' psychological states is critical to our research, we conducted an ablation study that validates the benefits of introducing the fusion layer. The removal of the fusion layer led to a performance drop of $\sim$ 11% and $\sim$ 13% for predicting climax and resolution, respectively. The loss of the protagonist's intent information impacts the climax prediction more, while the effect of emotion is reflected on resolution prediction. Compared to prior state-of-the-art methods, the average performance jumped by $\sim$ 20% higher success in modeling these elements of narrative structure in personal stories.

7.2 Limitations & Future Work

This section discusses the limitations of our current work and potential future work that can promote dedicated efforts taking forward the ideas presented in this dissertation. Below we enlist the limitations and future work related to each part of our work.

- Learning Knowledge-Enriched Social Event Representation:
  - Several recent works [Liu+20; Shw+20] have shown the importance of high-quality symbolic knowledge to generalize on commonsense information. However, we aggregate noisy commonsense knowledge data in our work. Though our models perform well in the current settings, finding ways to address the noise can help learn richer embeddings, thereby improving their utility in other tasks. Exploring better ways of mining web data and identifying cleaner resources can help handle this issue. The other way is to draw ideas from several recent studies on denoising [Lew+19; SJ19; Zho+19; SL21] and build robust models that can be resilient to a certain level of noise in data.
  - The social commonsense knowledge extracted using our method contains search results from random websites and could potentially have biases that humans propagate. Further, this data may not reflect socio-cultural dependencies associated with events. The motives and emotions related to specific actions and events could vary across cultures, geographical locations, and societies. In order to factor in those facets, we need to investigate methods [ATF20; For+20] to integrate such knowledge and weigh them from the perspective of inclusivity and applicability to data from those domains.

  - Though our current model works well on several downstream tasks, it is worth evaluating our model in a different domain like news stories to assess the adaptability and how much they apply to such domains comprising of rapidly changing interpretations of words depending on the values and ideologies of the event participants.

- Modeling Human Motives and Emotions from Personal Narratives:

  - As a part of tackling the challenges related to developing mentalizing models, we focus on the characters' motives and emotions in this work. However, several other dimensions are left unexplored, such as beliefs, thoughts, desires, goals, etc. []. This requires knowledge acquisition relevant to those dimensions and assessing the interplay between them for constructing better mental models of people.
  - Our work is based on many models primarily suited for short personal narratives. It might be a challenge to apply our models to long-format texts directly. Therefore, immediate future work is to explore ways to extend our model or transfer the knowledge from our model to longer text narratives.
  - The disadvantage of not accommodating knowledge related to socio-cultural norms is applicable for this work involving narratives because the underlying meaning of the story and how people react to them might be entirely dependent on the cultural context. As a critical limitation and the necessity to overcome this shortcoming, we highlight a growing concern towards building fair and inclusive machine learning applications.

- Modeling Narrative Structure in Short Personal Narratives:

- In this work, we primarily focus on two prominent elements of the narrative structure – climax and resolution. However, there is a scope to extend our work to identify fine-grained categories of narrative structure. There could be other types of stories that may not have a central conflict but have some inherent structure. Our work currently pays less attention to such types and the implicit structure present in them.
- Exploring the protagonist's psychological state is one of the several dimensions employed for modeling narrative structures. It will be interesting to delve deeper into modeling the relationships between the characters in the mental state space and identifying the causal patterns that lead to specific structures in narratives.
- Applicability to fictional narratives has not been investigated thoroughly in this work. Though it is beyond the scope of this work, we see it as a reasonable next step to understand how our mental state-enriched narrative structure models can detect structure in fictional narratives.

Besides these shortcomings, the data and modeling resources produced through this research have tremendous potential to be utilized in a wide variety of applications. These resources can be directly beneficial to many tasks such as narrative comprehension, retrieval, and generation [], empathetic dialogue generation, pattern analysis about human behavior in real-life phenomena like elections, cyberbullying, online hate speech, fake news, or propaganda analysis, to list a few.

7.3 Closing Remarks

In summary, our work sheds light on how we could leverage the mutual influence of narratives and mentalizing towards furthering the research in developing socially-aware AI systems. We have highlighted the strengths of our proposed computational methods and provided empirical evidence of their utility. Though we demonstrate improved performance on several tasks related to reasoning about mental states and narrative comprehension, we also expose the weaknesses of our research and identify several areas of improvement in each of the problems we solve. It is essential to expand on the ideas and resources produced through our research and recognize future investigations amenable to achieve the grand challenge of endowing social intelligence to machines. We hope that our work motivates a nuanced step in that direction and offers critical resources as well as useful insights for researchers embarking on the quest to build human-centric AI systems.

Bibliography

Abbott, H Porter (2020). *The Cambridge introduction to narrative*. Cambridge University Press.

Abrams, Meyer Howard and Geoffrey Harpham (2014). *A glossary of literary terms*. Nelson Education.

Acharya, Anurag, Kartik Talamadupula, and Mark A Finlayson (2020). "An atlas of cultural commonsense for machine reasoning". In: *arXiv preprint arXiv:2009.05664*.

Adams, Catherine et al. (2005). "Pragmatic language impairment: case studies of social and pragmatic language therapy". In: *Child Language Teaching and Therapy* 21.3, pp. 227–250.

Alexander, Franz, Bernard Trans Glueck, and Bertram D Lewin (1935). "The psychoanalysis of the total personality: The application of Freud's theory of the ego to the neuroses." In:

Alm, Cecilia Ovesdotter and Richard Sproat (2005). "Emotional sequencing and development in fairy tales". In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, pp. 668–674.

Anderson, Anne, Simon C Garrod, and Anthony J Sanford (1983). "The accessibility of pronominal antecedents as a function of episode shifts in narrative text". In: *Quarterly Journal of Experimental Psychology* 35.3, pp. 427–440.

Angeli, Gabor, Melvin Jose Johnson Premkumar, and Christopher D Manning (2015). "Leveraging linguistic structure for open domain information extraction". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 344–354.

Aram, Dorit and Sigalit Aviram (2009). "Mothers' storybook reading and kindergartners' socioemotional and literacy development". In: *Reading Psychology* 30.2, pp. 175–194.

Astington, Janet Wilde (2004). "Sometimes necessary, never sufficient: False-belief understanding and social competence". In: *Individual differences in theory of mind*. Psychology Press, pp. 24–49.

Astington, Janet Wilde and Janette Pelletier. "Theory of mind, language, and learning in the early years: Developmental origins of school readiness". In: *The development of social cognition and communication* (), pp. 205–230.

Badenes, Lid'on Villanueva, Rosa Ana Clemente Estevan, and Francisco J Garc'ıa Bacete (2000). "Theory of mind and peer rejection at school". In: *Social Development* 9.3, pp. 271–283.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473*.

Baker, Chris, Rebecca Saxe, and Joshua Tenenbaum (2011). "Bayesian theory of mind: Modeling joint belief-desire attribution". In: *Proceedings of the annual meeting of the cognitive science society*. Vol. 33. 33.

Baker, Chris L, Rebecca Saxe, and Joshua B Tenenbaum (2009). "Action understanding as inverse planning". In: *Cognition* 113.3, pp. 329–349.

Baker, Chris L et al. (2017). "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing". In: *Nature Human Behaviour* 1.4, pp. 1–10.

Baker, Collin F, Charles J Fillmore, and John B Lowe (1998). "The berkeley framenet project". In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pp. 86–90.

Baldassano, Christopher, Uri Hasson, and Kenneth A Norman (2018). "Representation of real-world event schemas during narrative perception". In: *Journal of Neuroscience* 38.45, pp. 9689–9699.

Balota, David A and Jennifer H Coane (2008). "Semantic memory". In:

Bamberg, Michael (2012). "Narrative analysis." In:

Bamman, David, Brendan OConnor, and Noah A Smith (2013). "Learning latent personas of film characters". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 352–361.

Barthes, Roland (1966). "Introduction to the structural analysis of the narrative". In:

Barthes, Roland and Lionel Duisit (1975). "An introduction to the structural analysis of narrative". In: *New literary history* 6.2, pp. 237–272.

Bauminger, Nirit (2007). "Brief report: Group social-multimodal intervention for HFASD". In: *Journal of autism and developmental disorders* 37.8, pp. 1605–1615.

Bauminger-Zviely, Nirit (2013). *Social and academic abilities in children with high-functioning autism spectrum disorders*.

Bauminger-Zviely, Nirit et al. (2013). "Increasing social engagement in children with high-functioning autism spectrum disorder using collaborative technologies in the school environment". In: *Autism* 17.3, pp. 317–339.

Beck, Julie (2015). "Lifes stories". In: *The Atlantic* 8.

Berman, Ruth A and Dan Isaac Slobin (2013). *Relating events in narrative: A crosslinguistic developmental study*. Psychology Press.

Boerma, Inouk E, Suzanne E Mol, and Jelle Jolles (2017). "The role of home literacy environment, mentalizing, expressive verbal ability, and print exposure in third and fourth graders reading comprehension". In: *Scientific Studies of Reading* 21.3, pp. 179–193.

Bosselut, Antoine and Yejin Choi (2019). "Dynamic knowledge graph construction for zero-shot commonsense question answering". In: *arXiv preprint arXiv:1911.03876*.

Bosselut, Antoine et al. (2017). "Simulating action dynamics with neural process networks". In: *arXiv preprint arXiv:1711.05313*.

Bosselut, Antoine et al. (2019). "COMET: Commonsense transformers for automatic knowledge graph construction". In: *arXiv preprint arXiv:1906.05317*.

Boyd, Ryan L, Kate G Blackburn, and James W Pennebaker (2020). "The narrative arc: Revealing core narrative structures through text analysis". In: *Science advances* 6.32, eaba2196.

Bridewell, Will and Alistair Isaac (2011). "Recognizing deception: A model of dynamic belief attribution". In: *2011 AAAI Fall Symposium Series*.

Brown, Steven (2020). "The Who System of the Human Brain: A System for Social Cognition About the Self and Others". In: *Frontiers in Human Neuroscience* 14, p. 224.

Brown, Steven et al. (2019). "Character mediation of story generation via protagonist insertion". In: *Journal of Cognitive Psychology* 31.3, pp. 326–342.

Bruner, Jerome (1990). *Acts of meaning.* Harvard university press.

— (1991a). "Self-making and world-making". In: *Journal of aesthetic education* 25.1, pp. 67–78.

— (1991b). "The narrative construction of reality". In: *Critical inquiry* 18.1, pp. 1–21.

Bruner, Jerome S (2009). *Actual minds, possible worlds*. Harvard university press.

Carnahan, Christina R, Pamela S Williamson, and Jennifer Christman (2011). "Linking cognition and literacy in students with autism spectrum disorder". In: *Teaching Exceptional Children* 43.6, pp. 54–62.

Carruthers, Peter and Peter K Smith (1996). *Theories of theories of mind*. Cambridge University Press.

Cer, Daniel et al. (2018). "Universal sentence encoder". In: *arXiv preprint arXiv:1803.11175*.

Ceran, Betul et al. (2012). "A semantic triplet based story classifier". In: *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, pp. 573–580.

Chafe, W. (1980). "The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production". In:

Chambers, Nathanael and Dan Jurafsky (2008). "Unsupervised learning of narrative event chains". In: *Proceedings of ACL-08: HLT*, pp. 789–797.

Charman, Tony and Yael Shmueli-Goetz (1998). "The relationship between theory of mind, language and narrative discourse: an experimental study." In: *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*.

Chaturvedi, Snigdha, Dan Goldwasser, and Hal Daume III (2016). "Ask, and shall you receive? understanding desire fulfillment in natural language text". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. 1.

Chen, Jiaao, Jianshu Chen, and Zhou Yu (2019). "Incorporating structured commonsense knowledge in story completion". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 6244–6251.

Chen, Qian et al. (2016). "Enhanced lstm for natural language inference". In: *arXiv preprint arXiv:1609.06038*.

Chu, Eric, Prashanth Vijayaraghavan, and Deb Roy (2018). "Learning Personas from Dialogue with Attentive Memory Networks". In: *arXiv preprint arXiv:1810.08717*.

Chung, Yu-An, Hung-Yi Lee, and James Glass (2017). "Supervised and unsupervised transfer learning for question answering". In: *arXiv preprint arXiv:1711.05345*.

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1, pp. 37–46.

Cutting, James E (2016). "Narrative theory and the dynamics of popular movies". In: *Psychonomic bulletin & review* 23.6, pp. 1713–1743.

Dawes, Robyn M (1999). "A message from psychologists to economists: mere predictability doesnt matter like it should (without a good story appended to it)". In: *Journal of Economic Behavior & Organization* 39.1, pp. 29–40.

De Fina, Anna (2008). "Who Tells Which Story and Why? Micro and Macros Contexts in Narrative". In:

De Fina, Anna and Alexandra Georgakopoulou (2011). *Analyzing narrative: Discourse and sociolinguistic perspectives*. Cambridge University Press.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Ding, Xiao et al. (2014). "Using structured events to predict stock price movement: An empirical investigation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1415–1425.

— (2015). "Deep learning for event-driven stock prediction". In: *Twenty-fourth international joint conference on artificial intelligence*.

— (2016). "Knowledge-driven event embedding for stock prediction". In: *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pp. 2133–2142.

Ding, Xiao et al. (2019). "Event Representation Learning Enhanced with External Commonsense Knowledge". In: *arXiv preprint arXiv:1909.05190*.

Dirkson, Anne, Suzan Verberne, and Wessel Kraaij (2019). "Narrative Detection in Online Patient Communities." In: *Text2Story@ ECIR*, pp. 21–28.

Dore, Rebecca A et al. (2018). "Theory of mind: A hidden factor in reading comprehension?" In: *Educational Psychology Review* 30.3, pp. 1067–1089.

Edunov, Sergey et al. (2018). "Understanding back-translation at scale". In: *arXiv preprint arXiv:1808.09381*.

Eisenberg, Joshua and Mark Finlayson (2017). "A simpler and more generalizable story detector using verb and character features". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2708–2715.

Eisenberg, Nancy (2014). *Altruistic emotion, cognition, and behavior (PLE: Emotion)*. Psychology Press.

Elson, David K (2012a). "Detecting story analogies from annotations of time, action and agency". In: *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative, Istanbul, Turkey*, pp. 91–99.

— (2012b). "Modeling narrative discourse". PhD thesis. Columbia University.

Ely, Jeffrey, Alexander Frankel, and Emir Kamenica (2015). "Suspense and surprise". In: *Journal of Political Economy* 123.1, pp. 215–260.

Fahlman, Scott E (2011). "Using Scone's multiple-context mechanism to emulate human-like reasoning". In: *2011 AAAI Fall Symposium Series*.

Felbo, Bjarke et al. (2017). "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm". In: *arXiv preprint arXiv:1708.00524*.

Feldman, Joshua, Joe Davison, and Alexander M Rush (2019). "Commonsense knowledge mining from pretrained models". In: *arXiv preprint arXiv:1909.00505*.

Fern'andez, Camila (2013). "Mindful storytellers: Emerging pragmatics and theory of mind development". In: *First Language* 33.1, pp. 20–46.

Ferstl, Evelyn C et al. (2008). "The extended language network: a meta-analysis of neuroimaging studies on text comprehension". In: *Human brain mapping* 29.5, pp. 581–593.

Fichman, Sveta et al. (2017). "Story grammar elements and causal relations in the narratives of Russian-Hebrew bilingual children with SLI and typical language development". In: *Journal of Communication Disorders* 69, pp. 72–93.

Finlayson, Mark Alan (2016). "Inferring Propp's functions from semantically annotated text". In: *The Journal of American Folklore* 129.511, pp. 55–77.

Finlayson, Mark Alan and Patrick Henry Winston (2006). "Analogical retrieval via intermediate features: The Goldilocks hypothesis". In:

Finlayson, Mark Mark Alan (2012). "Learning narrative structure from annotated folktales". PhD thesis. Massachusetts Institute of Technology.

Fishbach, Ayelet and Melissa J Ferguson (2007). "The goal construct in social psychology." In:

Forbes, Maxwell et al. (2020). "Social chemistry 101: Learning to reason about social and moral norms". In: *arXiv preprint arXiv:2011.00620*.

Freytag, Gustav (1894). *Die technik des dramas*. S. Hirzel.

Friedrich, Lynette Kohn and Aletha Huston Stein (1973). "Aggressive and prosocial television programs and the natural behavior of preschool children". In: *Monographs of the Society for Research in Child Development*, pp. 1–64.

Gallese, Vittorio and Hannah Wojciehowski (2011). "How stories make us feel: Toward an embodied narratology". In: *California Italian Studies* 2.1.

Gaonkar, Radhika et al. (2020). "Modeling label semantics for predicting emotional reactions". In: *arXiv preprint arXiv:2006.05489*.

Garc'ıa-P'erez, Rosa M, R Peter Hobson, and Anthony Lee (2008). "Narrative role-taking in autism". In: *Journal of Autism and Developmental Disorders* 38.1, pp. 156–168.

Gernsbacher, Morton Ann (2013). *Language comprehension as structure building*. Psychology Press.

Ghosh, Sayan et al. (2017). "Affect-lm: A neural language model for customizable affective text generation". In: *arXiv preprint arXiv:1704.06851*.

Goodwin, Travis et al. (2012). "UTDHLT: COPACETIC system for choosing plausible alternatives". In: * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 461–466.

Gordon, Andrew, Cosmin Bejan, and Kenji Sagae (2011). "Commonsense causal reasoning using millions of personal stories". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 25. 1.

Gordon, Andrew and Reid Swanson (2009). "Identifying personal stories in millions of weblog entries". In:

Gordon, Andrew S et al. (2013). "Identifying Personal Narratives in Chinese Weblog Posts." In:

Gould, Evelyn et al. (2011). "Teaching children with autism a basic component skill of perspective-taking". In: *Behavioral Interventions* 26.1, pp. 50–66.

Goyal, Amit, Ellen Riloff, and Hal Daum'e III (2010). "Automatically producing plot unit representations for narrative text". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 77–86.

Goyal, Amit, Ellen Riloff, and Hal Daum'e III (2013). "A computational model for plot units". In: *Computational Intelligence* 29.3, pp. 466–488.

Graesser, Arthur C, Brent Olde, and Bianca Klettke. "How does the mind construct and represent stories". In: ().

Granroth-Wilding, Mark and Stephen Clark (2016). "What happens next? event prediction using a compositional neural network model". In: *Thirtieth AAAI Conference on Artificial Intelligence*.

Gray, Heather M, Kurt Gray, and Daniel M Wegner (2007). "Dimensions of mind perception". In: *science* 315.5812, pp. 619–619.

Greenberg, Daniel L and Mieke Verfaellie (2010). "Interdependence of episodic and semantic memory: evidence from neuropsychology". In: *Journal of the International Neuropsychological Society: JINS* 16.5, p. 748.

Guajardo, Nicole R and Anne C Watson (2002). "Narrative discourse and theory of mind development". In: *The Journal of Genetic Psychology* 163.3, pp. 305–325.

Gui, Lin et al. (2017). "A question answering approach to emotion cause extraction". In: *arXiv preprint arXiv:1708.05482*.

Hakemulder, Jèmeljan (2000). *The moral laboratory: Experiments examining the effects of reading literature on social perception and moral self-concept*. Vol. 34. John Benjamins Publishing.

Happ'e, Francesca GE (1994). "An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults". In: *Journal of autism and Developmental disorders* 24.2, pp. 129–154.

Hardalov, Momchil, Ivan Koychev, and Preslav Nakov (2018). "Towards automated customer support". In: *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. Springer, pp. 48–59.

Harrison, Brent and Mark O Riedl (2016). "Towards learning from stories: An approach to interactive machine learning". In: *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Hearst, Marti A (1997). "Text Tiling: Segmenting text into multi-paragraph subtopic passages". In: *Computational linguistics* 23.1, pp. 33–64.

Henaff, Mikael et al. (2016). "Tracking the world state with recurrent entity networks". In: *arXiv preprint arXiv:1612.03969*.

Hermann, Karl Moritz et al. (2015). "Teaching machines to read and comprehend". In: *Advances in neural information processing systems*, pp. 1693–1701.

Hermans, Alexander, Lucas Beyer, and Bastian Leibe (2017). "In defense of the triplet loss for person re-identification". In: *arXiv preprint arXiv:1703.07737*.

Hogan, Patrick Colm (2003). *The mind and its stories: Narrative universals and human emotion*. Cambridge University Press.

Hopper, Robert and Rita C Naremore (1978). *Children's speech: A practical introduction to communication development*. HarperCollins Publishers.

Hughes, Claire (1998). "Finding your marbles: Does preschoolers' strategic behavior predict later understanding of mind?" In: *Developmental psychology* 34.6, p. 1326.

Hutto, Clayton and Eric Gilbert (2014). "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 8. 1.

Hutto, Daniel (2007). "The narrative practice hypothesis: origins and applications of folk psychology". In:

Hutto, Daniel D (2012). *Folk psychological narratives: The sociocultural basis of understanding reasons*. MIT press.

Imuta, Kana et al. (2016). "Theory of mind and prosocial behavior in childhood: A meta-analytic review." In: *Developmental psychology* 52.8, p. 1192.

Iyyer, Mohit et al. (2018). "Adversarial example generation with syntactically controlled paraphrase networks". In: *arXiv preprint arXiv:1804.06059*.

Jenkins, Jennifer M and Janet Wilde Astington (2000). "Theory of mind and social behavior: Causal models tested in a longitudinal study". In: *Merrill-Palmer Quarterly (1982-)*, pp. 203–220.

Jhala, Arnav (2008). "Exploiting structure and conventions of movie scripts for information retrieval and text mining". In: *Joint International Conference on Interactive Digital Storytelling*. Springer, pp. 210–213.

Ji, Yangfeng et al. (2017). "Dynamic entity representations in neural language models". In: *arXiv preprint arXiv:1708.00781*.

Jockers, Matthew L (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Johnson, Stephen B et al. (2008). "An electronic health record based on structured narrative". In: *Journal of the American Medical Informatics Association* 15.1, pp. 54–64.

Jorge, A et al. (2018). "First international workshop on narrative extraction from texts (Text2Story 2018)". In: *ECIR*, pp. 833–834.

Jorge, Al'ıpio M'ario et al. (2019). "The 2 nd International Workshop on Narrative Extraction from Text: Text2Story 2019". In: *European Conference on Information Retrieval*. Springer, pp. 389–393.

Kaland, Nils et al. (2002). "A newadvanced'test of theory of mind: evidence from children and adolescents with Asperger syndrome". In: *Journal of child psychology and psychiatry* 43.4, pp. 517–528.

Kimhi, Yael (2014). "Theory of mind abilities and deficits in autism spectrum disorders". In: *Topics in Language Disorders* 34.4, pp. 329–343.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kintsch, Walter and Edith Greene (1978). "The role of culture-specific schemata in the comprehension and recall of stories". In: *Discourse processes* 1.1, pp. 1–13.

Klin, Ami et al. (2002). "Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism". In: *Archives of general psychiatry* 59.9, pp. 809–816.

Kociskỳ, Tom'as et al. (2018). "The narrativeqa reading comprehension challenge". In: *Transactions of the Association for Computational Linguistics* 6, pp. 317–328.

Koopman, Bevan, Liam Cripwell, and Guido Zuccon (2017). "Generating clinical queries from patient narratives: a comparison between machines and humans". In: *Proceedings of the 40th international ACM SIGIR conference on Research and development in information retrieval*, pp. 853–856.

Labov, William (1972). *Language in the inner city: Studies in the Black English vernacular*. 3. University of Pennsylvania Press.

— (2001). *Uncovering the event structure of a narrative. Georgetown University Round Table*.

— (2006). "Narrative pre-construction". In: *Narrative inquiry* 16.1, pp. 37–45.

Labov, William and Joshua Waletzky (1997). "Narrative analysis: Oral versions of personal experience." In:

Lan, Wuwei et al. (2017). "A continuously growing dataset of sentential paraphrases". In: *arXiv preprint arXiv:1708.00391*.

Lanctot, Marc et al. (2017). "A unified game-theoretic approach to multiagent reinforcement learning". In: *Advances in Neural Information Processing Systems*, pp. 4190–4203.

Lehnert, Wendy G (1981). "Plot units and narrative summarization". In: *Cognitive science* 5.4, pp. 293–331.

Lenat, Douglas B et al. (1990). "Cyc: toward programs with common sense". In: *Communications of the ACM* 33.8, pp. 30–49.

Lengua, Liliana J (2003). "Associations among emotionality, self-regulation, adjustment problems, and positive adjustment in middle childhood". In: *Journal of Applied Developmental Psychology* 24.5, pp. 595–618.

Lerer, Adam and Alexander Peysakhovich (2018). "Learning social conventions in markov games". In: *arXiv preprint arXiv:1806.10071*.

Levi, Effi et al. (2020). "CompRes: A Dataset for Narrative Structure in News". In: *arXiv preprint arXiv:2007.04874*.

Lewis, Mike et al. (2019). "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension". In: *arXiv preprint arXiv:1910.13461*.

Li, Boyang et al. (2013). "Story generation with crowdsourced plot graphs". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 27. 1.

Li, Boyang et al. (2017). "Annotating high-level structures of short stories and personal anecdotes". In: *arXiv preprint arXiv:1710.06917*.

Li, Jiwei et al. (2016). "A persona-based neural conversation model". In: *arXiv preprint arXiv:1603.06155*.

Li, Zhongyang, Xiao Ding, and Ting Liu (2018a). "Constructing narrative event evolutionary graph for script event prediction". In: *arXiv preprint arXiv:1805.05081*.

— (2018b). "Generating reasonable and diversified story ending using sequence to sequence model with adversarial training". In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1033–1043.

— (2019). "Story ending prediction by transferable BERT". In: *arXiv preprint arXiv:1905.07504*.

Lind, Majse and Dorthe Kirkegaard Thomsen (2018). "Functions of personal and vicarious life stories: Identity and empathy". In: *Memory* 26.5, pp. 672–682.

Liu, Chia-Wei et al. (2016). "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation". In: *arXiv preprint arXiv:1603.08023*.

Liu, Hugo and Push Singh (2004). "ConceptNeta practical commonsense reasoning tool-kit". In: *BT technology journal* 22.4, pp. 211–226.

Liu, Ye et al. (2020). "Commonsense Evidence Generation and Injection in Reading Comprehension". In: *arXiv preprint arXiv:2005.05240*.

MacKay, Tommy, Fiona Knott, and Aline-Wendy Dunlop (2007). "Developing social interaction and understanding in individuals with autism spectrum disorder: A groupwork intervention". In: *Journal of Intellectual and Developmental Disability* 32.4, pp. 279–290.

Magliano, Joseph et al. (2012). "Aging and perceived event structure as a function of modality". In: *Aging, Neuropsychology, and Cognition* 19.1-2, pp. 264–282.

Magliano, Joseph P, Holly A Taylor, and Hyun-Jeong Joyce Kim (2005). "When goals collide: Monitoring the goals of multiple characters". In: *Memory & cognition* 33.8, pp. 1357–1367.

Mann, William C and Sandra A Thompson (1988). *Towards a functional theory of text organization*.

Manshadi, Mehdi, Reid Swanson, and Andrew S Gordon (2008). "Learning a Probabilistic Model of Event Sequences from Internet Weblog Stories." In: *FLAIRS Conference*, pp. 159–164.

Mar, Raymond A (2011). "The neural bases of social cognition and story comprehension". In: *Annual review of psychology* 62, pp. 103–134.

— (2018). "Evaluating whether stories can promote social cognition: Introducing the Social Processes and Content Entrained by Narrative (SPaCEN) framework". In: *Discourse Processes* 55.5-6, pp. 454–479.

Mar, Raymond A and Keith Oatley (2008). "The function of fiction is the abstraction and simulation of social experience". In: *Perspectives on psychological science* 3.3, pp. 173–192.

Mar, Raymond A, Jennifer L Tackett, and Chris Moore (2010). "Exposure to media and theory-of-mind development in preschoolers". In: *Cognitive Development* 25.1, pp. 69–78.

Mar, Raymond A et al. (2006). "Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds". In: *Journal of Research in Personality* 40.5, pp. 694–712.

Marchionini, Gary, Peter Liebscher, and Xia Lin (1991). "Authoring hyperdocuments: Designing for interaction". In: *Interfaces for Information Retrieval and Online Systems. Greenwood Press, New York, NY*, pp. 119–131.

Mason, Robert A and Marcel Adam Just (2009). "The role of the theory-of-mind cortical network in the comprehension of narratives". In: *Language and Linguistics Compass* 3.1, pp. 157–174.

McCabe, Allyssa, McCabe Allyssa, and Carole Peterson (1991). *Developing narrative structure*. Psychology Press.

McKeough, Anne (1992). "A neo-structural analysis of childrens narrative and its development". In: *The minds staircase: Exploring the conceptual underpinnings of childrens thought and knowledge*, pp. 171–188.

McNamara, Danielle S and Joe Magliano (2009). "Toward a comprehensive model of comprehension". In: *Psychology of learning and motivation* 51, pp. 297–384.

McQuillan, Martin (2000). "Introduction: Aporias of writing: Narrative and subjectivity". In: *The narrative reader*, pp. 1–34.

Modi, Ashutosh (2016). "Event embeddings for semantic script modeling". In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 75–83.

Mohammad, Saif (2013). "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales". In: *arXiv preprint arXiv:1309.5909*.

Mostafazadeh, Nasrin et al. (2017). "Lsdsem 2017 shared task: The story cloze test". In: *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pp. 46–51.

Mumper, Micah L and Richard J Gerrig (2017). "Leisure reading and social cognition: A meta-analysis." In: *Psychology of Aesthetics, Creativity, and the Arts* 11.1, p. 109.

Murray, Michael (2003). "Narrative psychology and narrative analysis." In:

Nguyen, Dai Quoc et al. (June 2018). "A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 327–333. DOI: 10.18653/v1/N18-2053. URL: https://www.aclweb.org/anthology/N18-2053.

Oatley, Keith (1995). "A taxonomy of the emotions of literary response and a theory of identification in fictional narrative". In: *Poetics* 23.1-2, pp. 53–74.

— (1999). "Why fiction may be twice as true as fact: Fiction as cognitive and emotional simulation". In: *Review of general psychology* 3.2, pp. 101–117.

Ouyang, Jessica and Kathleen McKeown (2015). "Modeling reportable events as turning points in narrative". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2149–2158.

Ouyang, Jessica and Kathy McKeown (2014). "Towards Automatic Detection of Narrative Structure." In: *LREC*, pp. 4624–4631.

ONeill, Brian and Mark Riedl (2011). "Toward a computational framework of suspense and dramatic arc". In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, pp. 246–255.

Palmer, Alan (2002). "The construction of fictional minds". In: *Narrative* 10.1, pp. 28–46.

— (2010). *Social minds in the novel*. The Ohio State University Press.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury (2005). "The proposition bank: An annotated corpus of semantic roles". In: *Computational linguistics* 31.1, pp. 71–106.

Papalampidi, Pinelopi, Frank Keller, and Mirella Lapata (2019). "Movie Plot Analysis via Turning Point Identification". In: *arXiv preprint arXiv:1908.10328*.

Papalampidi, Pinelopi et al. (2020). "Screenplay Summarization Using Latent Narrative Structure". In: *arXiv preprint arXiv:2004.12727*.

Parsons, Lauren et al. (2017). "A systematic review of pragmatic language interventions for children with autism spectrum disorder". In: *PloS one* 12.4, e0172242.

Paszke, Adam et al. (2019). "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems*, pp. 8026–8037.

Paul, Debjit and Anette Frank (2019). "Ranking and selecting multi-hop knowledge paths to better predict human needs". In: *arXiv preprint arXiv:1904.00676*.

Paynter, Jessica and Candida C Peterson (2013). "Further evidence of benefits of thought-bubble training for theory of mind development in children with autism spectrum disorders". In: *Research in Autism Spectrum Disorders* 7.2, pp. 344–348.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Perren, Sonja et al. (2007). "Pathways of behavioural and emotional symptoms in kindergarten children: What is the role of pro-social behaviour?" In: *European Child & Adolescent Psychiatry* 16.4, pp. 209–214.

Pichotta, Karl and Raymond Mooney (2014). "Statistical script learning with multi-argument events". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 220–229.

Pino, Maria Chiara and Monica Mazza (2016). "The use of literary fiction to promote mentalizing ability". In: *PloS one* 11.8, e0160254.

Piper, Andrew (2018). "Fictionality". In:

Polanyi, Livia (1981). "What stories can tell us about their teller's world". In: *Poetics Today* 2.2, pp. 97–112.

Polkinghorne, Donald E (1988). *Narrative knowing and the human sciences*. Suny Press.

Prasad, Rashmi et al. (2008). "The Penn Discourse TreeBank 2.0." In: *LREC*. Citeseer.

Prince, Gerald (2012). *A grammar of stories: An introduction*. Vol. 13. Walter de Gruyter.

Propp, Vladimir (2010). *Morphology of the Folktale*. Vol. 9. University of Texas Press.

Radford, Alec et al. (2018). "Improving language understanding by generative pre-training". In:

Rahimtoroghi, Elahe et al. (2014). "Minimal narrative annotation schemes and their applications". In: *7th Workshop on Intelligent Narrative Technologies*.

Rahimtoroghi, Elahe et al. (2017). "Modelling protagonist goals and desires in first-person narrative". In: *arXiv preprint arXiv:1708.09040*.

Rao, Anand S, Michael P Georgeff, et al. (1995). "BDI agents: from theory to practice." In: *Icmas*. Vol. 95, pp. 312–319.

Rashkin, Hannah et al. (2018a). "Event2mind: Commonsense inference on events, intents, and reactions". In: *arXiv preprint arXiv:1805.06939*.

Rashkin, Hannah et al. (2018b). "I know the feeling: Learning to converse with empathy". In:

Rashkin, Hannah et al. (2018c). "Modeling naive psychology of characters in simple commonsense stories". In: *arXiv preprint arXiv:1805.06533*.

Rashkin, Hannah et al. (2018d). "Towards empathetic open-domain conversation models: A new benchmark and dataset". In: *arXiv preprint arXiv:1811.00207*.

Riedl, Mark O and Robert Michael Young (2010). "Narrative planning: Balancing plot and character". In: *Journal of Artificial Intelligence Research* 39, pp. 217–268.

Rieffe, Carolien, Mark Meerum Terwogt, and Richard Cowan (2005). "Children's understanding of mental states as causes of emotions". In: *Infant and Child Development: An International Journal of Research and Practice* 14.3, pp. 259–272.

Riessman, Catherine Kohler (1993). *Narrative analysis*. Vol. 30. Sage.

Roemmele, Melissa, Cosmin Adrian Bejan, and Andrew S Gordon (2011). "Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning." In: *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pp. 90–95.

Rokach, Lior, Oded Maimon, and Mordechai Averbuch (2004). "Information retrieval system for medical narrative reports". In: *International Conference on Flexible Query Answering Systems*. Springer, pp. 217–228.

Rumelhart, David E (1975). "Notes on a schema for stories". In: *Representation and understanding*. Elsevier, pp. 211–236.

Ryan, Marie-Laure (1986). "Embedded narratives and tellability". In: *Style*, pp. 319–340.

— (1991). *Possible worlds, artificial intelligence, and narrative theory*. Indiana University Press.

Ryckman, Richard M (2004). "Theory of Personality". In: *USA. Michele Sordi*.

Sap, Maarten et al. (2019a). "Atomic: An atlas of machine commonsense for if-then reasoning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 3027–3035.

Sap, Maarten et al. (2019b). "SocialIQA: Commonsense Reasoning about Social Interactions". In: *arXiv preprint arXiv:1904.09728*.

Saxe, Rebecca and Nancy Kanwisher (2003). "People thinking about thinking people: the role of the temporo-parietal junction in theory of mind". In: *Neuroimage* 19.4, pp. 1835–1842.

Schafer, Stephen Brock (2016). *Exploring the Collective Unconscious in the Age of Digital Media*. IGI Global.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2015). "Improving neural machine translation models with monolingual data". In: *arXiv preprint arXiv:1511.06709*.

Serban, Iulian Vlad et al. (2017). "A hierarchical latent variable encoder-decoder model for generating dialogues". In: *Thirty-First AAAI Conference on Artificial Intelligence*.

Sergio, Gwenaelle Cunha and Minho Lee (2021). "Stacked DeBERT: All attention in incomplete data for text classification". In: *Neural Networks* 136, pp. 87–96.

Shamay-Tsoory, Simone G and Judith Aharon-Peretz (2007). "Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study". In: *Neuropsychologia* 45.13, pp. 3054–3067.

Shankar, Avi, Richard Elliott, and Christina Goulding (2001). "Understanding consumption: Contributions from a narrative perspective". In: *Journal of marketing Management* 17.3-4, pp. 429–453.

Shwartz, Vered et al. (2020). "Unsupervised commonsense question answering with self-talk". In: *arXiv preprint arXiv:2004.05483*.

Socher, Richard et al. (2013). "Reasoning with neural tensor networks for knowledge base completion". In: *Advances in neural information processing systems*, pp. 926–934.

Somasundaran, Swapna, Josef Ruppenhofer, and Janyce Wiebe (2008). "Discourse level opinion relations: An annotation study". In: *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pp. 129–137.

Souri, Alireza, Shafigheh Hosseinpour, and Amir Masoud Rahmani (2018). "Personality classification based on profiles of social networks users and the five-factor model of personality". In: *Human-centric Computing and Information Sciences* 8.1, p. 24.

Speer, Robyn, Joshua Chin, and Catherine Havasi (2017). "Conceptnet 5.5: An open multilingual graph of general knowledge". In: *Thirty-First AAAI Conference on Artificial Intelligence*.

Srivastava, Shashank, Snigdha Chaturvedi, and Tom Mitchell (2016). "Inferring interpersonal relations in narrative summaries". In: *Thirtieth AAAI Conference on Artificial Intelligence*.

Stanovsky, Gabriel et al. (2018). "Supervised open information extraction". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 885–895.

Stapleton, Karyn and John Wilson (2017). "Telling the story: Meaning making in a community narrative". In: *Journal of Pragmatics* 108, pp. 60–80.

Storm, William (2016). *Dramaturgy and Dramatic Character*. Cambridge University Press.

Suh, SY u and Tom Trabasso (1993). "Inferences during reading: Converging evidence from discourse analysis, talk-aloud protocols, and recognition priming". In: *Journal of memory and language* 32.3, pp. 279–300.

Sun, Yifu and Haoming Jiang (2019). "Contextual Text Denoising with Masked Language Models". In: *arXiv preprint arXiv:1910.14080*.

Sun, Yixing (2007). "Using the organizational and narrative thread structures in an e-book to support comprehension." PhD thesis.

Suway, Jenna G et al. (2012). "The relations among theory of mind, behavioral inhibition, and peer interactions in early childhood". In: *Social Development* 21.2, pp. 331–342.

Swanson, Reid et al. (2014). "Identifying narrative clause types in personal stories". In: *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 171–180.

Tager-Flusberg, Helen and Kate Sullivan (1995). "Attributing mental states to story characters: A comparison of narratives produced by autistic and mentally retarded individuals". In: *Applied Psycholinguistics* 16.3, pp. 241–256.

Tambwekar, Pradyumna et al. (2018). "Controllable Neural Story Plot Generation via Reinforcement Learning". In: *arXiv preprint arXiv:1809.10736*.

Tamir, Diana I and Mark A Thornton (2018). "Modeling the predictive social mind". In: *Trends in cognitive sciences* 22.3, pp. 201–212.

Tamir, Diana I et al. (2016a). "Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence". In: *Proceedings of the National Academy of Sciences* 113.1, pp. 194–199.

Tamir, Diana I et al. (2016b). "Reading fiction and reading minds: The role of simulation in the default network". In: *Social cognitive and affective neuroscience* 11.2, pp. 215–224.

Thorne, Avril (2004). "Putting the person into social identity". In: *Human Development* 47.6, pp. 361–365.

Thorne, Avril, Neill Korobov, and Elizabeth M Morgan (2007). "Channeling identity: A study of storytelling in conversations between introverted and extraverted friends". In: *Journal of research in personality* 41.5, pp. 1008–1031.

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

Trabasso, Tom, Paul Van den Broek, and So Young Suh (1989). "Logical necessity and transitivity of causal relations in stories". In: *Discourse processes* 12.1, pp. 1–25.

Trabasso, Tom and Paul Van Den Broek (1985). "Causal thinking and the representation of narrative events". In: *Journal of memory and language* 24.5, pp. 612–630.

Urooj, Aisha, Amir Mazaheri, Mubarak Shah, et al. (2020). "MMFT-BERT: Multimodal Fusion Transformer with BERT Encodings for Visual Question Answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 4648–4660.

Valls-Vargas, Josep, Jichen Zhu, and Santiago Ontañ'on (2015). "Narrative hermeneutic circle: Improving character role identification from natural language text via feedback loops". In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems*, pp. 5998–6008.

Vijayaraghavan, Prashanth, Eric Chu, and Deb Roy (2020). "DAPPER: Learning Domain-Adapted Persona Representation Using Pretrained BERT and External Memory". In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 643–652.

Vijayaraghavan, Prashanth and Deb Roy (2021a). "Lifelong Knowledge-Enriched Social Event Representation Learning". In: *16th European Chapter of the Association for Computational Linguistics (EACL 2021)*.

— (2021b). "Modeling Human Motives and Emotions from Personal Narratives Using External Knowledge And Entity Tracking". In: *Proceedings of The Web Conference 2021*.

Weber, Noah, Niranjan Balasubramanian, and Nathanael Chambers (2018). "Event representations with tensor-based compositions". In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

Wellman, Henry M and Joan G Miller (2008). "Including deontic reasoning as fundamental to theory of mind". In: *Human Development* 51.2, pp. 105–135.

Wieting, John and Kevin Gimpel (2017). "Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations". In: *arXiv preprint arXiv:1711.05732*.

Wilmot, David and Frank Keller (2020). "Suspense in short stories is predicted by uncertainty reduction over neural story representation". In: *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 1763–1788.

Winston, Patrick Henry (2014). *The genesis story understanding and story telling system a 21st century step toward artificial intelligence*. Tech. rep. Center for Brains, Minds and Machines (CBMM).

Wood, Barbara S (1976). "Children and communication: Verbal and nonverbal language development." In:

Wyer Jr, Robert S (2014). *Knowledge and Memory: The Real Story: Advances in Social Cognition, Volume VIII*. Psychology Press.

Xia, Yingce et al. (2017). "Deliberation networks: Sequence generation beyond one-pass decoding". In: *Advances in Neural Information Processing Systems*, pp. 1784–1794.

Xiong, Hao et al. (2019). "Modeling coherence for discourse neural machine translation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 7338–7345.

Yang, Zichao et al. (2016). "Hierarchical attention networks for document classification". In: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480–1489.

Yoshida, Wako, Ray J Dolan, and Karl J Friston (2008). "Game theory of mind". In: *PLoS computational biology* 4.12.

Yuan, Xingdi et al. (2017). "Machine comprehension by text-to-text neural question generation". In: *arXiv preprint arXiv:1705.02012*.

Zhang, Zhengyan et al. (2019). "ERNIE: Enhanced language representation with informative entities". In: *arXiv preprint arXiv:1905.07129*.

Zhou, Shuyan et al. (2019). "Improving robustness of neural machine translation with multi-task learning". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 565–571.

Zunshine, Lisa (2006). *Why we read fiction: Theory of mind and the novel*. Ohio State University Press.

Zwaan, Rolf A (2004). "The immersed experiencer: Toward an embodied theory of language comprehension". In: *Psychology of learning and motivation* 44, pp. 35–62.

— (2016). "Situation models, mental simulations, and abstract concepts in discourse comprehension". In: *Psychonomic bulletin & review* 23.4, pp. 1028–1034.

Zwaan, Rolf A et al. (1998). "Constructing multidimensional situation models during reading". In: *Scientific studies of reading* 2.3, pp. 199–220.