M-sense: Modeling Narrative Structure in Short Personal Narratives Using Protagonist's Mental Representations

Prashanth Vijayaraghavan, Deb Roy
MIT Media Lab, 75 Amherst Street, Cambridge, MA, 02139 USA
pralav@mit.edu, dkroy@media.mit.edu

Abstract

Narrative is a ubiquitous component of human communication. Understanding its structure plays a critical role in a wide variety of applications, ranging from simple comparative analyses to enhanced narrative retrieval, comprehension, or reasoning capabilities. Prior research in narratology has highlighted the importance of studying the links between cognitive and linguistic aspects of narratives for effective comprehension. This interdependence is related to the textual semantics and mental language in narratives, referring to characters' motivations, feelings or emotions, and beliefs. However, this interdependence is hardly explored for modeling narratives. In this work, we propose the task of automatically detecting prominent elements of the narrative structure by analyzing the role of characters' inferred mental state along with linguistic information at the syntactic and semantic levels. We introduce a STORIES dataset of short personal narratives containing manual annotations of key elements of narrative structure, specifically climax and resolution. To this end, we implement a computational model that leverages the protagonist's mental state information obtained from a pre-trained model trained on social commonsense knowledge and integrates their representations with contextual semantic embed-dings using a multi-feature fusion approach. Evaluating against prior zero-shot and supervised baselines, we find that our model is able to achieve significant improvements in the task of identifying climax and resolution.

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 (Beck 2015). Humans share their personal experiences by picking specific events or facts and weaving them together to make meaning. These are referred to as personal narratives, a form of autobiographical storytelling that gives shape to experiences. Polkinghorne (1988) 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 the 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.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

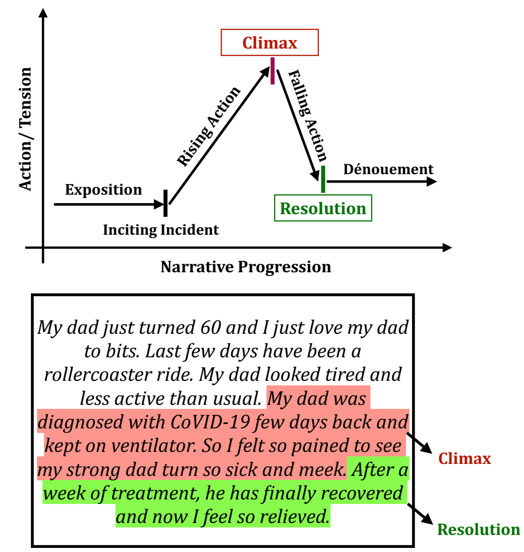

Figure 1: (Top) Freytag's Pyramid. (Bottom) Highlights of climax and resolution for a sample personal narrative.

Several recent works (Dore et al. 2018; Yuan et al. 2017; Chung, Lee, and Glass 2017; Kočiskỳ et al. 2018; Mostafazadeh et al. 2017) 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 (McCabe, Allyssa, and Peterson 1991; Finlayson and Winston 2006; Marchionini, Liebscher, and Lin 1991), (b) linking related stories that form a narrative thread towards theme generation (Berman and Slobin 2013), (c) summarization of stories (Lehnert 1981; Papalampidi et al. 2020) and (d) story ending prediction or generation (Chen, Chen, and Yu 2019; Li et al. 2013; Mostafazadeh et al. 2017), (e) commonsense reasoning (Goodwin et al. 2012; Gordon, Bejan, and Sagae 2011), to list a few. Several narrative theories have been proposed such as Freytag (Freytag 1894), Prince (Prince 2012), Bruner (Bruner 1991, 2009), Labov & Waletzky (Labov and Waletzky 1997), 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' (Bruner 1991) could correspond to (a) Freytag's 'climax' – referring to the 'turning point' of the fortunes of the protagonist (Abrams and Harpham 2014) 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 (Labov and Waletzky 1997; Labov 2006). 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, while the other structural elements are more likely to occur in longer narratives. Figure 1 (Left) shows Freytag's pyramid containing the key elements of the narrative structure and Figure 1 (right) contains highlights of climax and resolution for a sample personal narrative. Thus, our work aims to develop computational approaches using aspects of information retrieval, NLP, and psychology, and model the key elements of narrative structure – MRE & resolution. As an operating definition, we consider an MRE to be contained in a sentence(s) based on the following criteria – it is an explicit event that can be reported as the summary of the story and occurs at the highest tension point of the story. Similarly, an event qualifies as 'resolution' if it usually occurs after the MRE and resolves the dramatic tension in the narrative.

Recently, Papalampidi et al. (Papalampidi, Keller, and Lapata 2019) 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 et al. 2017) and automatically extracting them from free text. Ouyang et al. (Ouyang and McKeown 2015)'s study on predicting MRE in narratives is the closest work to the problem considered in this paper. While most of these methods rely on syntactic, semantic, surface-level affect, or narrative features obtained using hand-engineering or pre-trained semantic embedding methods to model narrative structure, we investigate the role of a 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 a basis for this study in prior theoretical frameworks (Murray 2003; Ryan 1986; Ouyang and McKeown 2014; Lehnert 1981; Schafer 2016) 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 (Vijayaraghavan and Roy 2021) that learns to embed characters' mental states using an external memory module. Our contributions are summarized below:

- A STORIES[1] corpus containing 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 via multi-feature fusion.
- 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.

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 interpersonal relationships and emotion trajectories, identifying enemies, allies, heroes; story-level analysis – story representation, predicting story endings, modeling story suspense, and creative or artistic storytelling, to list a few.

Several studies have analyzed the literature in narratology and formulated different goals and annotation labels associated with narratives for modeling their structure. Elson's (Elson 2012) 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 (Rao, Georgeff et al. 1995) and intention-based story planning (Riedl and Young 2010). Previous studies like (Gordon, Bejan, and Sagae 2011) have analyzed personal weblog stories containing everyday situations. Rahimtoroghi et al. (Rahimtoroghi et al. 2014) and Swanson et al. (Swanson et al. 2014) 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 grammar in story understanding, As explained earlier, (Papalampidi, Keller, and Lapata 2019)'s dataset for analyzing turning points is a valuable addition in this area of work. Moreover, there have been consistent efforts (Jorge et al. 2018, 2019) that study the link between Information Retrieval (IR) and narrative representations from a given text. These include works that exploit narrative structure

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

[2]Short for **M**ental **S**tate **E**nriched **N**arrative **S**tructure mod**E**l

in movies for IR (Jhala 2008), detect and retrieve narratives in health domain (patient communities & medical reports) (Johnson et al. 2008; Rokach, Maimon, and Averbuch 2004; Dirkson, Verberne, and Kraaij 2019; Koopman, Cripwell, and Zuccon 2017), identify narrative structures in news stories (Boyd, Blackburn, and Pennebaker 2020; Levi et al. 2020) or generate summaries from screenplays or novels (Papalampidi et al. 2020), to name a few. Given this wide spectrum of work, we leverage mental state representation models that are pre-trained 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 to our task.

Dataset Collection

First, we collect Reddit posts from two communities: /r/offmychest and /r/confession using the PushShift API [3]. Next, we filter the collected data to retain only those posts that do not contain tags like "[Deleted]", "NSFW" [4] or "over_18". Finally, we further narrow down the aggregated posts using a BERT-based story classifier. The data collection pipeline is described in the supplementary material[5].

Annotation

Our STORIES dataset comprises $63k$ Reddit Personal Narratives, of which $\sim 2.3k$ were annotated. The annotated subset contains a total of $\sim 5.1k$ climax sentences and $\sim 4.5k$ resolution sentences. The annotation process is explained below.

Setup We created a user interface for MTurk workers to make the annotation procedure convenient for capturing the key elements of the narrative structure. The MTurk workers can highlight parts of the text that qualify as climax and resolution using red and green colors respectively. Three annotators were involved in the annotation process. Each worker is presented a sampled text from the 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.

Agreements We measure the inter-annotator agreement (IAA) at the sentence-level. For sentence-level agreement, the Fleiss's kappa (κ) (Fleiss and Cohen 1973) is 0.646 and 0.756 for climax and resolution respectively. We also compute the mean annotation distance ($\mathcal{D}$), i.e., the distance between two annotations for each category, normalized by story length (Papalampidi, Keller, and Lapata 2019). For the annotated data, we find that: $\mathcal{D}_{climax} = 1.764$ and $\mathcal{D}_{resolution} = 1.59$.

[3] https://pushshift.io/

[4] NSFW – not safe for work

[5] Refer supplementary material: http://arxiv.org/abs/2302.09418

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 2. While the average position for climax (0.61) coincides with the peak, we observe that the resolution contents occur later in the story. We observe that substantial agreement is achieved for both the climax and resolution. Clearly, we obtain higher agreement values for resolution than the climax. Figure 3 displays sample annotations (e.g. multi-sentence or non-contiguous highlights) from our STORIES dataset.

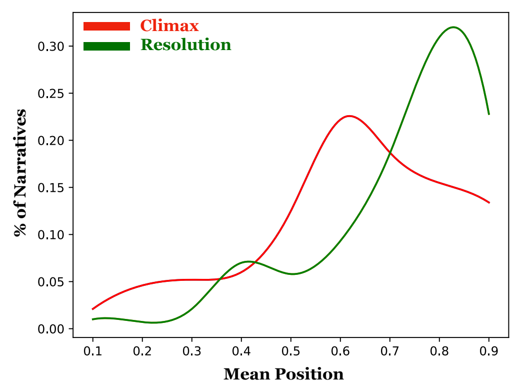

Figure 2: Distributions of the positions of Climax & Resolution.

M-sense: Modeling Narrative Structure

In this paper, 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. 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. Beyond linguistic features extracted from narratives, we focus on a dominant aspect in which a narrative is formed or presented, that is an account of characters' mental states – motives and emotions. Thus, we leverage transfer learning from pretrained models trained to infer characters' mental states from a narrative. We implement 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. Our M-sense model consists of the following components:

Ensemble Sentence Encoders, which computes per-sentence linguistic & mental state embeddings.

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

Story Encoder, which maps the fused encodings into a sequence of bidirectionally contextualized embeddings.

Interaction layer, which estimates state transition across sequential context windows to identify the boundaries.
Classification layer, which involves linear layers to eventually calculate the label probabilities.

Ensemble Sentence Encoders

In this work, we aim to exploit both linguistic and mental state features for an enhanced model for narratives.

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.In our M-sense model, we use a token-level BERT-based sentence encoder[5]. Here, each sentence is prepended with a special $[CLS]$ token at the beginning of each sentence and appended with a $[SEP]$ token at the end of each sentence in the narrative. We apply both position and segment embeddings and feed to the pre-trained BERT model as:

$$H = [h^1_{[CLS]}, .., h^1_{N_1}, h^1_{[SEP]}, .., h^i_{[CLS]}, .., h^i_{N_i}, .., h^L_{[SEP]}] = \text{BERT}(T) \quad (1)$$

The hidden representation of the 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 $h^i_{[CLS]}$ and denote the output semantic embeddings as: $H^{xSem}_{sents} = [h^1, h^2, ..., h^L]$.

Incorporating Protagonist's Mental Representation Prior studies have established how the progression of a story 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 (Palmer 2002; Mohammad 2013; Alm and Sproat 2005). We follow a recent work (Vijayaraghavan and Roy 2021) that implements a NEMO 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 3 contains the overview of the NEMO architecture. 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. 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, $\text{STORYENTENC}(\cdot)$, 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:

$$(\hat{H}^i_{xIntent}, \tilde{H}^i_{xReact}) = \text{STORYENTENC}(S_i, S_{<i}, e_j, m); \quad \forall m \in \{xIntent, xReact\} \quad (2)$$

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: $H^{xintent}_{sents} = [\hat{h}^1, .., \hat{h}^L]$ and $H^{xReact}_{sents} = [\tilde{h}^1, .., \tilde{h}^L]$ respectively.

Transformer-Based Fusion Layer

Given multiple sentence-level embeddings, we apply a fusion strategy to derive 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 reactions $(xReact)$ of the i^{th} sentence respectively. Drawing ideas from the literature of multimodal analysis (Urooj et al. 2020), we concatenate the multiple latent vectors and introduce a special token $[FUSE]$[6] 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: $h^i_{fuse} = \text{TF}(\|^K_{k=0}\ h^{ik})$, where TF refers to the transformer encoder layer and h^{i0} (i.e. when $k = 0$) is set to the trainable $[FUSE]$ vector.

Story Encoder

We apply Transformer layers on the top of the sentence embeddings, referred to as *Inter-sentence Transformer*, to extract narrative-level features. Intuitively, the transformer layer focuses on different sentences in the narrative, and produces context-aware sentence embeddings. This is given as:

$$\begin{aligned} \hat{H}^l &= LayerNorm(\hat{C}^{l-1} + \text{MHA}(\hat{C}^{l-1})) \\ \hat{C}^l &= LayerNorm(\hat{H}^l + \text{FFL}(\hat{H}^l)) \\ C_{sents} &= [c^1, c^2, ..., c^L] = \hat{C}^{n_L} \end{aligned} \quad (3)$$

where $\hat{C}^0 = PE(H_{sents})$, PE refers to the positional encoding, $LayerNorm$ refers to layer normalization operation, MHA is the multi-head attention operation and FFL is a feed-forward layer (Vaswani et al. 2017). 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}.

Interaction Layer

In this layer, we compute the transition of state across sentences by measuring similarity metrics in the embedding space between sequential context windows and concatenating them with contextualized embeddings. 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. Finally, we get the interaction-feature enhanced context-aware embeddings: $E_{sents} = [e^1, e^2, ..., e^L]$.

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

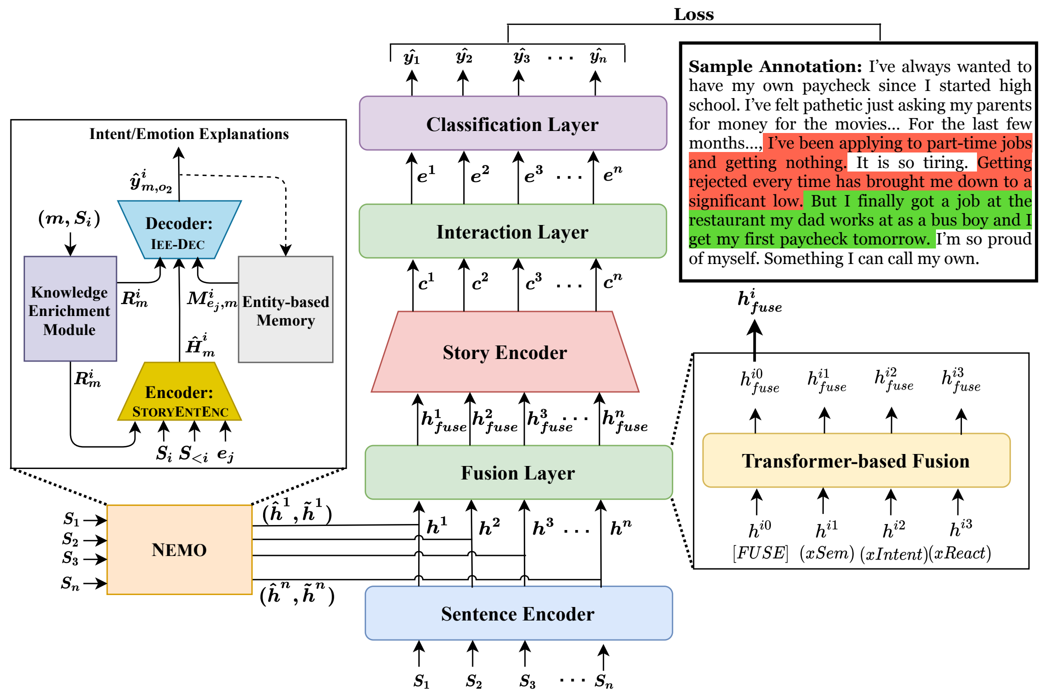

Figure 3: Illustration of our M-sense model. Note that $h^{i1} = h^i; h^{i2} = \hat{h}^i; h^{i3} = \tilde{h}^i$ relate to semantics, intents and emotions of the i^{th} sentence respectively. (Top-right) Sample annotation of climax (Red) & Resolution (Green) by one of the annotators.

Classification Layer

The resulting embeddings are mapped to a C-dimensional output using a softmax-based classification layer (here, $C = 3$). This step is given as: $\hat{y}_i = softmax(f_s(e^i))$.

Experiments

We run experiments to study the following questions:
RQ1: How does our model compare with other baselines for identifying climax and resolution in short narratives?
RQ2: How do various model components contribute to the overall performance? To what extent do mental state representations play a role for our classification task?

Overall Predictive Performance (RQ1)

Baselines We compare our model with a set of carefully selected zero-shot & supervised baselines[5], 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 earlier.

Heuristic baseline, which labels the sentence that is the closest semantic neighbour of the post title as climax, the last sentence in the narrative as the resolution.

A recent work (Wilmot and Keller 2020) has explored surprise a measure of suspense in narratives. (Ely, Frankel, and Kamenica 2015)'s surprise is defined as the amount of change from previous sentence to the current sentence in the narrative[5]. We encode the sentences in the story using the following approaches and eventually compute suspense measures for our classification task.

GloVeSim (Pennington, Socher, and Manning 2014), **BERT** (Devlin et al. 2018), **USE** (Cer et al. 2018), which computes semantic embeddings using average word vectors (using GloVe) or Transformer-based models.

STORYENC (Papalampidi, Keller, and Lapata 2019), which uses the hierarchical RNN based language model to encode sentences in the story.

STORYENTENC (Vijayaraghavan and Roy 2021), which encodes the sentences in the story from the protagonist's perspective. Here, we denote intent and emotion embeddings as $(E_{int} = H^{xIntent}_{sents})$ and $(E_{emo} = H^{xReact}_{sents})$ respectively.

CAM and **TAM** (Papalampidi, Keller, and Lapata 2019) consist of bidirectional LSTM model with the latter model having an additional interaction layer to compute boundaries between the topics in each story.

M-sense–Fusion, which is a variant of our M-sense model without mental state embeddings.

M-sense, which is our complete model incorporating protagonist's mental representation.

Results Table 1 outlines the results of our evaluation. We report the performance of simple baselines, of which the distribution baseline turns out to be the strongest. The heuristic baseline performs slightly better than the random baseline. This suggests that the Reddit post title contains relevant signal to predict the climax while 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

Models	$F_1 \uparrow$		$D \downarrow$	
	C	**R**	**C**	**R**
Random	0.196	0.143	29.05	30.57
Distribution	0.274	0.315	**15.79**	**14.42**
Heuristic	0.217	0.147	23.74	26.82
GloVeSim	0.312	0.344	12.06	11.65
BERT$_{tok}$	0.408	0.441	9.37	8.09
BERT$_{sent}$	0.352	0.366	10.88	9.73
USE$_{sent}$	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 ($\pm$0.022)	0.609 ($\pm$0.008)	5.90 ($\pm$3.18)	5.02 ($\pm$2.82)
CAM	0.578 ($\pm$0.019)	0.604 ($\pm$0.0032)	6.58 ($\pm$4.02)	5.44 ($\pm$3.05)
M-sense	**0.694*** ($\pm$0.0027)	**0.743*** ($\pm$0.0015)	**4.15** ($\pm$1.84)	**3.20** ($\pm$1.06)

Table 1: Evaluation Results with F_1 score per class & percent D reported for each model. $\uparrow, \downarrow$ indicate if higher/ lower values mean better performance respectively. * refers to significance ($p < 0.05$) over TAM using a paired T-Test.

the simple baselines in terms of both the evaluation metrics. As expected, sentence-level BERT/USE performs worse 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 exhibit an overall enhanced classification capability. We find that the intent embedding (E_{int}) helps achieve the best zero-shot performance for detecting climax. A 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 (Papalampidi, Keller, and Lapata 2019) applied for similar tasks. As we can see, supervised fine-tuning approaches easily beat the earlier results obtained using zero-shot methods. Finally, our M-sense model achieves an absolute improvement of $\sim 20.07\%$ and $\sim 22\%$ for climax and resolution prediction respectively.

Ablation Study (RQ2)

To evaluate the contributions of each component 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 2 shows the results of our study. We modify one component at a time and report their performance using F_1 metric. This involves either replacing a

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 2: We report F_1 score per class with non-default modeling choices for each component of our model.

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 our 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.

Influence of Mental State Embeddings: In this study, we examine the necessity of a fusion layer and probe the influence of protagonist's mental state embeddings towards our classification task. Notably, the results in Table 2 validate the benefits of the fusion layer and demonstrate the relative performance gains obtained with intent and emotion embeddings. Without the fusion layer, we observe a performance drop of $\sim 11\%$ and $\sim 13\%$ for predicting climax and resolution respectively. The loss of the protagonist's intent features impacts the climax prediction more. This is analogous to the effect emotion features has on resolution prediction.

Analysis and Discussion

Effect of Story Length: Here, we compare the performance of different sentence encoders with and without fusion layer for detecting climax in narratives with varying length. Figure 4 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 stable and relatively better outcomes for longer narratives (story length > 14). With the introduction of mental state representation through fusion layer, the F_1 score improved significantly irrespective of the sentence encoder used.

Error Analysis: To estimate why our model augmented with mental state embeddings performs better, we conduct error analysis between 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

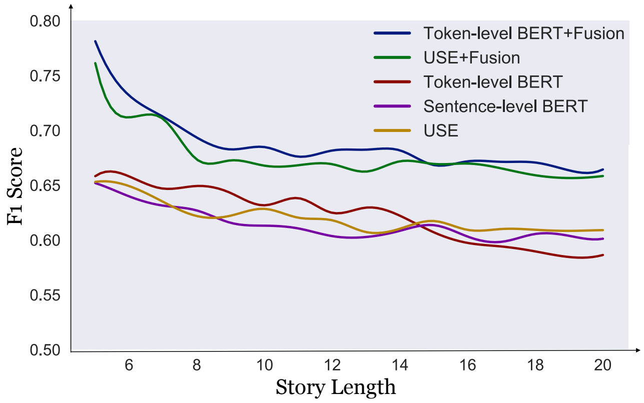

Figure 4: Performance of sentence encoders for detecting climax in story with varying length.

out of the following analysis: (a) Using VADER[7] (Hutto and Gilbert 2014) a weighted composite sentiment score is computed for each sentence in the narrative, and (b) Using state classification (Rashkin et al. 2018; Vijayaraghavan and Roy 2021), we assess Maslow's motivation categories associated with sentences predicted as climax/resolution in the narrative and analyze for any error patterns emerging from it. For resolution, the M-sense−Fusion makes $\sim$ 28% more mistakes than M-sense model for narratives with homogeneous endings (i.e. narratives having same sentiment sentences in the neighbourhood of resolution closer to the story's end). M-sense−Fusion model is unable to discern clearly and predicts a different sentence as resolution. Based on our analysis (b), we find that M-sense gains significantly over the M-sense−Fusion when the ground-truth climax sentences belong to "Esteem" and "Love/Belonging" categories[5].

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. (Papalampidi, Keller, and Lapata 2019) introduced a TRIPOD dataset containing a corpus of movie synopses annotated with turning points (TPs). 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 away. By their definitions for each of the TP categories (Papalampidi, Keller, and Lapata 2019), 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. We use the cast information collected from IMDb as a part of this dataset and evaluate our model's performance on this out-of-domain dataset. We first apply our M-sense trained on our STORIES corpus directly and evaluate its zero-shot performance (referred as ZS). We assume the top character from the IMDb cast section to be the movie protagonist. Though this may not always be true, it measures how our model fares on this dataset

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

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

Table 3: Results of evaluation on TRIPOD dataset. We report (%) D results for TP4 & TP5 relevant to this work.

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. We then fine-tune our model with the TRIPOD's training set. This is denoted by M-sense(FT).

Results

We display our model's performance compared to the best performing TAM reported in the original work (Papalampidi, Keller, and Lapata 2019). 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 to each TP. Similarly, TAM+Entities enriched the model with entity information by applying co-reference resolution and obtain entity-specific representations. Table 3 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.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 distance 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 M-sense model can predict climax and resolution in stories beyond just personal narratives, albeit limited by story length at this point.

Conclusion

Towards modeling high-level narrative structure, we construct a dataset of personal narratives from Reddit containing annotations of climax and resolution sentences. Next, we introduce a deep neural model, referred to as M-sense, that learns to effectively integrate 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 protagonist's mental state embeddings. Our model is able to achieve $\sim$ 20% higher success in prediction task 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 AI tools that can interact with users though stories.

References

Abrams, M. H.; and Harpham, G. 2014. *A glossary of literary terms*. Nelson Education.

Alm, C. O.; and Sproat, R. 2005. Emotional sequencing and development in fairy tales. In *International Conference on Affective Computing and Intelligent Interaction*, 668–674. Springer.

Beck, J. 2015. Life's stories. *The Atlantic*, (8).

Berman, R. A.; and Slobin, D. I. 2013. *Relating events in narrative: A crosslinguistic developmental study*. Psychology Press.

Boyd, R. L.; Blackburn, K. G.; and Pennebaker, J. W. 2020. The narrative arc: Revealing core narrative structures through text analysis. *Science advances*, 6(32): eaba2196.

Bruner, J. 1991. The narrative construction of reality. *Critical inquiry*, 18(1): 1–21.

Bruner, J. S. 2009. *Actual minds, possible worlds*. Harvard university press.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Chen, J.; Chen, J.; and Yu, Z. 2019. Incorporating structured commonsense knowledge in story completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6244–6251.

Chung, Y.-A.; Lee, H.-Y.; and Glass, J. 2017. Supervised and unsupervised transfer learning for question answering. *arXiv preprint arXiv:1711.05345*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dirkson, A.; Verberne, S.; and Kraaij, W. 2019. Narrative Detection in Online Patient Communities. In *Text2Story@ ECIR*, 21–28.

Dore, R. A.; Amendum, S. J.; Golinkoff, R. M.; and Hirsh-Pasek, K. 2018. Theory of mind: A hidden factor in reading comprehension? *Educational Psychology Review*, 30(3): 1067–1089.

Elson, D. K. 2012. Detecting story analogies from annotations of time, action and agency. In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative, Istanbul, Turkey*, 91–99.

Ely, J.; Frankel, A.; and Kamenica, E. 2015. Suspense and surprise. *Journal of Political Economy*, 123(1): 215–260.

Finlayson, M. A.; and Winston, P. H. 2006. Analogical retrieval via intermediate features: The Goldilocks hypothesis. (MIT CSAIL Technical Report No. MIT-CSAIL-TR-2006-071).

Fleiss, J. L.; and Cohen, J. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3): 613–619.

Freytag, G. 1894. *Die technik des dramas*. S. Hirzel.

Goodwin, T.; Rink, B.; Roberts, K.; and Harabagiu, S. 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)*, 461–466.

Gordon, A.; Bejan, C.; and Sagae, K. 2011. Commonsense causal reasoning using millions of personal stories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25.

Hutto, C.; and Gilbert, E. 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*, volume 8.

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

Johnson, S. B.; Bakken, S.; Dine, D.; Hyun, S.; Mendonça, E.; Morrison, F.; Bright, T.; Van Vleck, T.; Wrenn, J.; and Stetson, P. 2008. An electronic health record based on structured narrative. *Journal of the American Medical Informatics Association*, 15(1): 54–64.

Jorge, A.; Campos, R.; Jatowt, A.; and Nunes, S. 2018. First international workshop on narrative extraction from texts (Text2Story 2018). In *ECIR*, 833–834.

Jorge, A. M.; Campos, R.; Jatowt, A.; and Bhatia, S. 2019. The 2 nd International Workshop on Narrative Extraction from Text: Text2Story 2019. In *European Conference on Information Retrieval*, 389–393. Springer.

Kočiskỳ, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K. M.; Melis, G.; and Grefenstette, E. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6: 317–328.

Koopman, B.; Cripwell, L.; and Zuccon, G. 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*, 853–856.

Labov, W. 2006. Narrative pre-construction. *Narrative inquiry*, 16(1): 37–45.

Labov, W.; and Waletzky, J. 1997. Narrative analysis: Oral versions of personal experience. *Journal of Narrative & Life History*, 7: 3–38.

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

Levi, E.; Mor, G.; Shenhav, S.; and Sheafer, T. 2020. CompRes: A Dataset for Narrative Structure in News. *arXiv preprint arXiv:2007.04874*.

Li, B.; Cardier, B.; Wang, T.; and Metze, F. 2017. Annotating high-level structures of short stories and personal anecdotes. *arXiv preprint arXiv:1710.06917*.

Li, B.; Lee-Urban, S.; Johnston, G.; and Riedl, M. 2013. Story generation with crowdsourced plot graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27.

Marchionini, G.; Liebscher, P.; and Lin, X. 1991. Authoring hyperdocuments: Designing for interaction. *Interfaces for Information Retrieval and Online Systems. Greenwood Press, New York, NY*, 119–131.

McCabe, A.; Allyssa, M.; and Peterson, C. 1991. *Developing narrative structure*. Psychology Press.

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

Mostafazadeh, N.; Roth, M.; Louis, A.; Chambers, N.; and Allen, J. 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*, 46–51.

Murray, M. 2003. *Narrative psychology and narrative analysis.*, 95–112. Qualitative research in psychology: Expanding perspectives in methodology and design. Washington, DC, US: American Psychological Association. ISBN 1-55798-979-6 (Hardcover).

Ouyang, J.; and McKeown, K. 2014. Towards Automatic Detection of Narrative Structure. In *LREC*, 4624–4631.

Ouyang, J.; and McKeown, K. 2015. Modeling reportable events as turning points in narrative. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2149–2158.

Palmer, A. 2002. The construction of fictional minds. *Narrative*, 10(1): 28–46.

Papalampidi, P.; Keller, F.; Frermann, L.; and Lapata, M. 2020. Screenplay Summarization Using Latent Narrative Structure. *arXiv preprint arXiv:2004.12727*.

Papalampidi, P.; Keller, F.; and Lapata, M. 2019. Movie Plot Analysis via Turning Point Identification. *arXiv preprint arXiv:1908.10328*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Polkinghorne, D. E. 1988. *Narrative knowing and the human sciences*. Suny Press.

Prince, G. 2012. *A grammar of stories: An introduction*, volume 13. Walter de Gruyter.

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

Rao, A. S.; Georgeff, M. P.; et al. 1995. BDI agents: from theory to practice. In *Icmas*, volume 95, 312–319.

Rashkin, H.; Bosselut, A.; Sap, M.; Knight, K.; and Choi, Y. 2018. Modeling naive psychology of characters in simple commonsense stories. *arXiv preprint arXiv:1805.06533*.

Riedl, M. O.; and Young, R. M. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39: 217–268.

Rokach, L.; Maimon, O.; and Averbuch, M. 2004. Information retrieval system for medical narrative reports. In *International Conference on Flexible Query Answering Systems*, 217–228. Springer.

Ryan, M.-L. 1986. Embedded narratives and tellability. *Style*, 319–340.

Schafer, S. B. 2016. *Exploring the Collective Unconscious in the Age of Digital Media*. IGI Global.

Swanson, R.; Rahimtoroghi, E.; Corcoran, T.; and Walker, M. 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)*, 171–180.

Urooj, A.; Mazaheri, A.; Shah, M.; 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*, 4648–4660.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vijayaraghavan, P.; and Roy, D. 2021. Modeling Human Motives and Emotions from Personal Narratives Using External Knowledge And Entity Tracking. *Proceedings of The Web Conference 2021*.

Wilmot, D.; and Keller, F. 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*, 1763–1788.

Yuan, X.; Wang, T.; Gulcehre, C.; Sordoni, A.; Bachman, P.; Subramanian, S.; Zhang, S.; and Trischler, A. 2017. Machine comprehension by text-to-text neural question generation. *arXiv preprint arXiv:1705.02012*.