"where is this relationship going?": Understanding Relationship Trajectories in Narrative Text

Keen You and **Dan Goldwasser**
Department of Computer Science, Purdue University
{you54, dgoldwas}@purdue.edu

Abstract

We examine a new commonsense reasoning task: given a narrative describing a social interaction that centers on two protagonists, systems make inferences about the underlying relationship trajectory. Specifically, we propose two evaluation tasks: Relationship Outlook Prediction MCQ and Resolution Prediction MCQ. In Relationship Outlook Prediction, a system maps an interaction to a relationship outlook that captures how the interaction is expected to change the relationship. In Resolution Prediction, a system attributes a given relationship outlook to a particular resolution that explains the outcome. These two tasks parallel two real-life questions that people frequently ponder upon as they navigate different social situations: *"where is this relationship going?"* and *"how did we end up here?"*. To facilitate the investigation of human social relationships through these two tasks, we construct a new dataset, ***Social Narrative Tree***, which consists of 1250 stories documenting a variety of daily social interactions. The narratives encode a multitude of social elements that interweave to give rise to rich commonsense knowledge of how relationships evolve with respect to social interactions. We establish baseline performances using language models and the accuracies are significantly lower than human performance. The results demonstrate that models need to look beyond syntactic and semantic signals to comprehend complex human relationships.

1 Introduction

A relationship between two people is constantly being shaped by their social interactions (Duck, 1994). For example, if two people have a proper conversation after a heated argument, they may become more intimate as they understand each other better. On the other hand, they may insist on their own views and part ways. For humans, being able to reason about relationship trajectories is crucial in achieving personal goals while avoiding conflicts with other people. This skill is formally termed as *social competence*, which humans easily acquire (Rubin et al., 1995).

In contrast to humans' innate ability to perceive social situations, machines struggle in developing social understanding. For example, machines' performance in reasoning about motivation and emotional reactions given a short context is significantly lower compared to human performance (Sap et al., 2019). Moreover, identifying the intents and reactions of characters in narratives poses another challenge for machines (Goyal et al., 2010; Chaturvedi et al., 2016; Rahimtoroghi et al., 2017; Rashkin et al., 2018a). Reasoning about relationship trajectories adds yet another layer of difficulty as in addition to the challenge of identifying the implicit factors that guide characters' behavior, such as intents, reactions and mental states, systems also need to understand how these elements collectively contribute to the evolution of relationships. This additional layer of understanding is challenging primarily because the impact of a particular event on a relationship is unique to each social scenario. For instance, an argument can help resolve differences or deepen them, depending on the personalities, intents and reactions of the characters involved in it.

The resources and methods discussed in current work do not directly address these challenges. A fixed set of mappings from events to relationship impact, in a style similar to Event2Mind (Rashkin et al., 2018b), is not sufficient to capture the uniqueness of each social situation. Similarly, SocialIQA (Sap et al., 2019) inspects various elements individually but does not unify them into a single force that influences a relationship. Other current datasets

*Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 168–178
Barcelona, Spain (Online), December 12–13, 2020

(Mostafazadeh et al., 2016; Zellers et al., 2018, 2019) cover a broad spectrum of commonsense knowledge, but do not focus on the social aspect. Although it can be annotated for such inspection such as in the investigation of intents and reactions (Rashkin et al., 2018a), this method is not pertaining to relationship analyses because a meaningful change may not always be present. For example, here is a description of a social scenario sampled from the ROCStories corpus: *Tina decided on going hiking with her friend Tony. They both decided on a difficult path. Upon ascension, Tina fell and cut her leg. They both decided it was too dangerous for them to continue. Tony carried Tina to the car and decided on mini-golf instead.* The story provides a few facts regarding a social interaction but does not provide sufficient details to draw conclusions about how the relationship between the two protagonists is changed because of the interaction. Unanswered questions include *"did Tina appreciate Tony's gesture?"* and *"did they bond as a result?"*. All these deficiencies in using current resources and methods to study human relationships call for the need of new resources that encode implicit relationship trajectories that are driven by the interweaving of social elements.

In this work, we introduce ***Social Narrative Tree***[1], a corpus of 1250 social narratives documenting a variety of social interactions, each centers on two protagonists. It is built incrementally from ten seeds in five narrative stages and the story branches into five different paths at each stage. This effectively captures the different possibilities which a relationship trajectory can take on at each diverging point. We choose the narrative stages – seed, buildup, climax, resolution and outlook - based on previous narrative analyses (Freytag, 1896; Prince, 1973; Labov, 1997), ensuring that intensity varies at different points of the narrative which simulates fluctuating intensity levels in real-life social situations and thus provides a natural space for relationships to develop in. Using ***Social Narrative Tree***, we set up two evaluation tasks, Relationship Outlook Prediction MCQ and Resolution Prediction MCQ. In each task, the text corresponding to Relationship Outlook or Resolution is removed from the stories for prediction based on other stages. The branching of each story at different stages creates a natural notion of similarity among stories, which can be measured by the number of stages two stories share. This facilitates the setup of confounding choices as the wrong candidates are not completely irrelevant in terms of textual information. This forces systems to truly understand the underlying relationship trajectories instead of relying solely on language modelling. The best accuracy achieved by our BERT-based model (Devlin et al., 2019) (˜60%) is significantly lower than human performance (˜80%). These experimental results demonstrate that models need to look beyond syntactic and semantic signals to comprehend complex human relationships.

In summary, our contributions are as follows: (1) we introduce ***Social Narrative Tree***, a corpus of social narratives with contextualized social elements contributing to relationship trajectories. (2) we introduce new evaluation frameworks using Relationship Outlook Prediction and Resolution Prediction, with meaningful confounding candidate choices that force systems not to depend only on textual information. (3) we establish baseline performances using language models and justify the importance of looking beyond textual information in understanding human relationships.

2 Related Work

Manually constructed scripts are used to represent structured knowledge in 1970-80s (Schank and Abelson, 1977). Subsequently, narrative event chains, unsupervised generation of such representations are introduced. Narrative event chains are sequences of events revolving around one central protagonist (Chambers and Jurafsky, 2008). Stemming from narrative event chains is the narrative cloze task where one event chain is removed and systems are required to fill in the blanks (Chambers and Jurafsky, 2008). The task is further refined to multiple choice form - Multiple Choice Narrative Cloze (MCNC) where different choices are randomly sampled from events that do not belong to the chain (Granroth-Wilding and Clark, 2016). Swaf Af (Zellers et al., 2018) and Hella Swag (Zellers et al., 2019) are datasets for next event prediction, containing multiple choice questions covering a wide range of grounded situations that are constructed from video captions. Furthermore, (Mostafazadeh et al., 2016) create ROCStories, a corpus of 50k commonsense stories each of five-sentence long and propose Story Cloze Test where systems are required to select the most plausible

[1]The dataset is publicly available at https://github.com/karenacorn99/Social-Narrative-Tree

ending for an incomplete narrative. In our work, Relationship Outlook Prediction and Resolution Prediction are multiple-choice tasks of similar motivation with a focus on social implications.

Various social elements are extensively studied in many works. (Sap et al., 2018) present ATOMIC, a knowledge graph of commonsense knowledge with 877k descriptions in free text form and focuses on 'if-then' relationships of causes, effects and attributes. Automatic construction of such knowledge base is explored in Comet (Bosselut et al., 2019). With ATOMIC as the foundation, SocialIQA, a dataset of 38,000 multiple-choice questions about social situations is constructed (Sap et al., 2019). Event2Mind is a corpus of phrasal verbs that is constructed to support the examination of the intents and reactions of common situations (Rashkin et al., 2018b). (Mohammad, 2018) presents NRC VAD Lexicon, a corpus of 20,000 English words with human ratings of valence, arousal and dominance while NRC Affect Intensity Lexicon (AIL) provides emotional categories and associated real values for approximately 6,000 English words (Mohammad, 2017). SocialSent (Hamilton et al., 2016) assigns tokens sentiment scores with contexts taken into consideration. One set of lexicons is constructed from the subreddit r/relationships which we use in our error analyses. Other works model relationships, between literary characters (Iyyer et al., 2016) or countries (Han et al., 2019), described in text using an unsupervised neural model, by mapping them to a latent space. Elements of psychology in ROCstories are analyzed in (Rashkin et al., 2018a) where ROCStories are annotated with motivations and emotional reactions of the characters involved. We create ***Social Narrative Tree*** with various social and psychological elements embedded in each story and these elements together contribute to the rise and fall of a relationship.

3 Collaborative Construction of Social Narratives

Our main contribution in this paper is the construction of a social narrative corpus consisting of short stories each describing the evolution of a relationship between two characters. The corpus is designed to capture how different social behaviors result in different relationship outcomes. To accomplish this goal we follow a fixed narrative structure consisting of five stages: seed (*exposition*), buildup (*rising action*), climax, resolution and outlook (*denouement*). The stages are based on "Freytag's Pyramid" (Freytag, 1896) and other more recent work analyzing repeating narrative structures (Prince, 1973; Labov, 1997). Fundamentally, a minimal story consists of an initial state, a final state and an intermediate state that transitions the initial state to the final state (Prince, 1973). More specifically, the intermediate state can be further broken down into individual stages to make a story more appealing and informative. Particularly, we adopt the phases of orientation clause, Most Reportable Event (MRE) and resolution (Labov, 1997) to make up the intermediate state, corresponding to the buildup, climax and resolution stages in our framework. The relationship in each story develops as the intensity of the narrative fluctuates.

The dataset is constructed in a collaborative way to help capture the impact of different behaviors at each narrative stage. We use Mechanical Turk, and provide annotators with a partial story created independently, and ask them to complete the next stage. Each MTurker is limited to provide three responses in total at each stage with all three responses belonging to different prompts.

3.1 Crowdsourcing Framework

Seed. The seed is a one-sentence description that introduces two protagonists, their initial relationship and a general social scenario that they are going to be involved in. The dataset is built on 10 initial seeds, which are created from the first sentences of 10 randomly selected stories from the ROCStories dataset (Mostafazadeh et al., 2016) by adding names and removing collective nouns. We inspect and reselect the randomly sampled seeds to ensure that there are no repetitions in social scenarios. This helps to elicit a variety of social behaviors and associated relationship impact. Among the 10 pairs of protagonists, seven pairs have the same gender, two pairs have different genders and one pair includes the name "*Sam*" which MTurkers have different gender interpretations. All the seeds have the same initial relationship description, *"friend"*, but the social scenario that the two protagonists are involved in provides additional information on the intimacy level between the two people (at one's home versus a public place). Each seed contains a predicate connecting the two protagonists. Among these predicates, *"asked"* and *"invited"* are active, *"receive"* is passive while *"realized"*, *"were"* and

"went to" have a neutral voice.

Buildup. We instruct MTurkers to provide a one-sentence continuation of the seed, which provides further information on the relationship between the protagonists or on the social scenario. Five different responses are collected for each seed. We specifically state in the instruction that only information relevant to the relationship of the protagonists should be given. This ensures that the plot is compact and the development of the relationship is the driving force of the narrative. We first programmatically filter responses that are completely irrelevant or do not meet the length requirement. We then manually filter all the remaining responses and re-annotate those that do not meet the requirements until the desired number of buildups are collected for each seed. We allow the addition of characters other than the two main protagonists in the responses, diversifying the types of events present in the dataset.

Climax. We create 50 partial narratives, consisting of the 10 seeds and their associated buildups, which serve as the input for the collection of climax events. We prompt MTurkers to provide a two-sentence continuation for each seed and buildup combination as the climax event, which we define as an event that has a significant impact on the relationship between the people involved. Five responses meeting the requirement are collected for each partial narrative, forming 250 seed-buildup-climax combinations.

Resolution and Outlook. We prompt MTurkers to provide a two-sentence continuation of the partial story as the resolution and a one-sentence continuation of the resolution as the relationship outlook. The purpose of resolution is to resolve, either successfully or not, the conflict brought up in the partial narratives. Relationship outlook states the effect of the complete interaction described in the narrative on the relationship between protagonists. Five resolution-outlook pairs are collected for each one of the 250 partial narratives, capturing how different choices of resolving conflicts can impact relationships. At the end of this stage, 1250 complete social narratives are created.

3.2 Dataset Overview

We illustrate the overall tree structure of Social Narrative Tree in Fig. 1 and present some basic statistics of the dataset in Tab. 1. In Tab. 2, we summarize the most frequent five predicates for each narrative stage. In this analysis, we use NLTK (Bird et al., 2009) to extract and lemmatize verbs and discard a list of stop words including *be*, *will*, *do* and *can*. The most used verbs in each stage are consistent with the general purpose of each stage. The seed initiates a social interaction *(invite, ask, receive)* and in buildups, the desires *(want, ask, decide)* and mental states *(think, feel)* of the protagonists are revealed. Climax is where the actual relationship-changing event happens *(start, go, get)* while resolution and outlook contain a mixture of further actions *(say, get, make)*, mental states *(feel)* and a notion of change *(decide, become, realize)*.

3.3 Relationship Trajectories in *Social Narrative Tree*

We carry out various exploratory analyses on ***Social Narrative Tree*** to examine the relationship trends present in the dataset with respect to social elements. The results validate that the stories contain a broad range of possible ways that relationships can unfold under different circumstances. In addition, the relationship trajectories present in the stories also exhibit some associations with different social elements, verifying that valuable social behavior commonsense knowledge is embedded in the narratives.

Sentiment Polarity. We associate a sentiment polarity label with each stage using TextBlob (Loria et al., 2014). The sentiment polarity distribution for each stage is shown in Tab. 3. The sentiment polarity distribution for climax is fairly equal (115 pos vs 106 neg) but positive resolutions and outlooks significantly outnumber negative resolutions and outlooks (678 pos vs 377 neg, 650 pos vs 269 neg). This indicates that in ***Social Narrative Tree***, social interactions are more likely to project positively. We further inspect overall relationship trends by breaking down stories into different combinations of polarities across stages and present the results in Tab. 4 and 5. From Tab. 4, we see that both positive and negative buildup branch equally into positive and negative climax (350 pos vs 350 neg, 150 pos vs 145 neg), giving both trends opportunities to develop and reducing the chance of monotonous plots. From resolution to outlook in Tab. 5, however, both positive and negative resolution are more likely to lead to positive outcomes. This observation could potentially tie back to people's general goal of avoiding conflicts with other people (Rubin et al., 1995) which MTurkers instill into the stories.

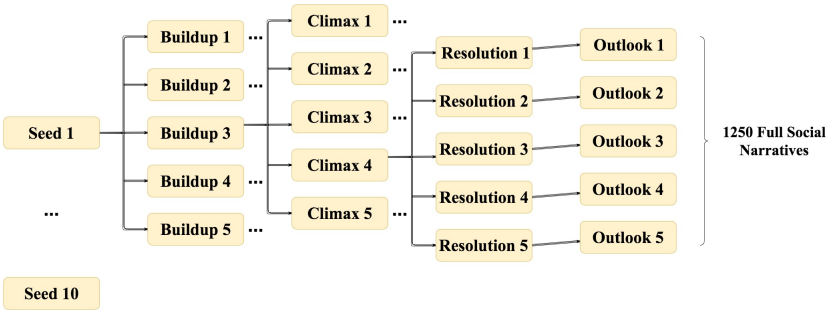

Figure 1: Visualization of Social Narrative Tree's overall structure.

	Seed	Buildup	Climax	Resolution	Outlook	Full Story
# sentences	1	1	2	2	1	7
# uni. instances	10	50	250	1250	1238	1250
# vocab	70	427	1583	3366	2294	4474
avg token #	12.5	24.7	35.2	30.7	19.1	122.2
max token #	15	40	92	108	57	211
min token #	10	14	12	6	3	66

Table 1: Social Narrative Tree basic information.

seed	buildup	climax	resolution	outlook
invite	want	start	decide	decide
ask	think	go	feel	become
go	ask	get	tell	feel
receive	decide	want	say	realize
realize	feel	say	get	make

Table 2: Most frequent five verbs for each stage.

	seed	buildup	climax	resolution	outlook
positive	0	32	115	678	650
negative	0	13	106	377	269
neutral	10	5	29	195	331

Table 3: Sentiment polarity distribution at each narrative stage.

buildup	climax	count
pos	pos	350
pos	neg	350
neg	pos	150
neg	neg	145

Table 4: Sentiment polarity trends between buildup and climax.

res	outlook	count
pos	pos	403
pos	neg	123
neg	pos	163
neg	neg	105

Table 5: Sentiment polarity trends between resolution and outlook.

Valence, Arousal, Dominance (VAD). For each stage we compute a vector of length three as the VAD representation (Mohammad, 2018). Within a stage, we retrieve the VAD scores for all tokens and take the maximum value for each dimension to represent the entire stage. We further associate VAD scores with general relationship trajectories represented using sentiment polarity by inspecting the VAD scores in different subsets of the corpus and present the results in Tab. 6 and 7. In Tab. 6, positive outlook is associated with higher valence and dominance in climax, resolution and outlook. It is also associated with lower arousal in climax and resolution. In Tab 7, positive climax is associated with lower arousal in climax but higher arousal in resolution. These results indicate that in ***Social Narrative Tree***, relationship trajectories are closely related to VAD which is an important social behavior indicator. This close relation can be transformed to rich commonsense knowledge if studied in-depth.

Affect Intensity. Affect Intensity assigns a real value to a lexicon in one of the four dimensions – joy, fear, anger and sadness (Mohammad, 2017). We assign a binary vector of length four for each stage for every story, indicating whether that particular dimension is present in the text for that stage. We analyze how different affect dimensions relate to general relationship trajectories by dividing the stories into two subsets using climax sentiment polarity. In each subset, we further separate the stories into two groups, one with positive trend and the other with negative trend – positive trend means the story lands on a positive relationship outlook and vice versa. Within each group, we compute the percentage of resolutions that contain words from each Affect dimension. Fig. 2 and 3 display the results for the positive climax set and the negative climax set respectively. In Fig. 2, resolutions in a positive trend contain a higher proportion of joy and a lower proportion of fear, anger and sadness. The distribution of Affect dimensions is similar in Fig. 3, except resolutions following a negative climax with a positive trend has a higher proportion of fear. Given a positive or negative relationship-changing event, this analysis relates mental states, another

outlook	climax			resolution			outlook		
	V	A	D	V	A	D	V	A	D
positive	0.8796	0.7537	0.7189	0.8900	0.7372	0.7349	0.9123	0.6721	0.7409
negative	0.8697	0.7627	0.7099	0.8631	0.7503	0.7195	0.8395	0.6656	0.6797

Table 6: VAD scores of climax, resolution and outlook partitioned by outlook sentiment polarity.

climax	climax			resolution			outlook		
	V	A	D	V	A	D	V	A	D
positive	0.8973	0.7612	0.7513	0.8870	0.7424	0.7292	0.8827	0.6559	0.7188
negative	0.8617	0.7723	0.6911	0.8630	0.7253	0.7237	0.8752	0.6502	0.7034

Table 7: VAD scores of climax, resolution and outlook partitioned by climax sentiment polarity.

important social behavior indicator, in resolution to relationship trajectories after the event, which is an important aspect of social behavior knowledge.

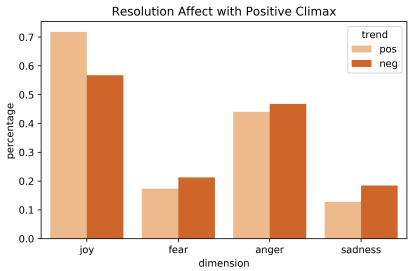

Figure 2: Resolution Affect dimensions in positive climax set.

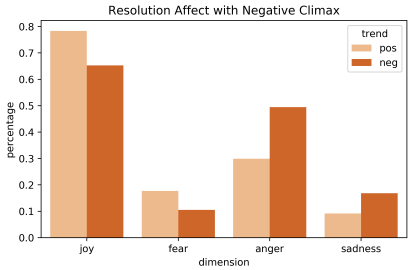

Figure 3: Resolution Affect dimensions in negative climax set.

4 Task Formulation

4.1 Relationship Outlook Prediction: *"Where is this relationship going?"*

In this task, Relationship Outlook is removed from stories and systems need to fill in the blank by choosing the most reasonable outlook from a pool of five candidates. Solving this task requires an understanding of social norms, particularly, what is expected of a relationship after a sequence of events happen in a particular social context. A sample question is shown below and choice D is the correct answer. In Choice A, the emotion of embarrassment is correctly inferred but "felt more comfortable" is inconsistent with the fact that Neil left the celebration. Choice B and Choice E are irrelevant to the given context. In choice C, "not realize that the event was not a casual one" is a correct statement with respect to the context but the expectation of them remaining friends is a socially incorrect deduction.

Neil went to a celebration with his Vietnamese friend James. Neil was a little nervous as he had never been to an ethnic celebration before, however, James reassured him that it was going to be very casual. When Neil arrived, the event was anything but casual. As he looked around the room, he saw women in dresses and men in suits and he was wearing t-shirt and jeans! James insisted that what Neil was wearing was fine, but Neil still disagreed. Neil got angry at James for lying to him and then left the celebration in tears.

What is the most likely relationship outlook?

(A) *Although at first Neil was embarrassed, throughout the night he felt more comfortable with James, and his confidence in his friend grew exponentially.*

(B) *James and Neil realized how much they had in common, and it strengthened their friendship.*

(C) *They remained friends because Neil had not realise that the event was not a casual one.*

(D) ***Neil felt like James could no longer be trusted after lying to him about the celebration, so he thought it would be best to stop talking to James.***

(E) *They will remain friends and more than likely to place more bets on sports or challenges that come between each other, it's friendly competition of two friends' egos.*

4.2 Resolution Prediction: *"How did we end up here?"*

After Resolution is removed from stories, systems need to reason about the cause of a relationship outlook by selecting one of the five resolution candidates. To select the correct resolution, systems need to understand the necessary conditions that

will result in a particular relationship outcome. An example Resolution Prediction question is given below. Choice B is correct as it completes the narrative both logically and socially correct. Choices A and D are relevant, but a romantic relationship would not have started if her feelings remain hidden. Choice C diverts from the context by mentioning a gift exchange whereas choice E is unlikely before the beginning of a romantic relationship.

Matthew asked his friend Emma to go for a walk in the park. The weather was really fine so they both took their dogs to the park. Emma was so pleased to be asked by Matthew to spend time with him at the park. Emma had a secret crush on Matthew and wondered if he felt the same but she was too scared to bring up the subject and potentially ruin their friendship. _____. This confession of feelings led to the beginning of a romantic relationship.

Which resolution best fills the blank?

(A) *She never said anything about that before. It was a mystery to him.*

(B) ***Matthew could tell from how Emma was acting that she liked him. He turned to face her and told her how he felt.***

(C) *Emma was very moved by Matthew's decision to give back the book, and agreed to accept it back. Emma let Matthew know how grateful his gesture was.*

(D) *She wanted to ask but didn't. He was clueless.*

(E) *Emma and Matthew held each others hands, kissed, and decided then that they wanted to spend the rest of their lives together.*

5 Experiments

5.1 Evaluation Tasks Setup

For each of the two evaluation tasks, we have two context setups, varying in the amount of information provided for systems to make predictions. In Outlook Prediction, *partial context* consists of only the resolution whereas *full context* consists of all of the remaining stages – seed, buildup, climax and resolution. Similarly, in Resolution Prediction, *partial context* consists of only the outlook whereas *full context* consists of all of the remaining stages – seed, buildup, climax and outlook. We will refer to *partial context* and *full context* as *context* below, indicating the corresponding narrative stages for the particular evaluation task and context setting at concern.

We create four sets of 1250 multiple-choice questions for each context setup for each task. For each story, we pair with the correct answer four confounding choices. The first confounding choice comes from one of the four stories that share the same climax as the current story. Namely, the two stories differ only in resolution and outlook. The second confounding choice comes from one of the 24 stories that share the same buildup and the third confounding choice comes from the context of one of the 124 stories that share the same seed. The final confounding choice is from one of the 1125 stories that stem from a different seed. Below, we will refer to the confounding choices as conf-diff-seed, conf-same-seed, conf-buildup and conf-climax to notate the source of a particular confounding choice where conf-buildup means the choice comes from a story that shares the same buildup, etc. Each confounding choice is randomly picked from their respective candidate pool. The names in confounding choices that are picked from other seeds are replaced by the names of current characters. The tree structure of dataset provides a natural notion of similarity among stories, thus, the confounding choices should be of different levels of difficulties, with conf-climax being the most difficult, followed by conf-buildup, conf-same-seed and with conf-diff-seed being the easiest.

We perform 10-fold cross validation on each of the four 1250 multiple-choice question sets. Each fold consists of the 125 stories built from the same seed and we run three random restarts for each fold, taking the best predicting accuracy on the validation fold as the performance for that fold. We use the average accuracy across the 10 folds as the criterion to select hyper-parameters and report the best accuracy for each model in Tab. 8. All of our BERT-based models use BERT-base with 110M parameters, with HuggingFace's PyTorch implementation (Wolf et al., 2019).

5.2 Baseline Models

Average word2vec. We represent both the context and each answer choice using the averaged word2vec representation of each individual word making up the sentences. We then compute the cosine similarity between each context-candidate choice pair. The answer choice that results in the highest cosine similarity value with the context is selected as the prediction.

Pretrained Bert For Next Sentence Prediction. We use BertForNextSentencePrediction to solve the tasks where each context and each answer choice is treated as a single sequence. Each context-answer pair is assigned a single score for entailment and we evaluate the softmax of the 5 scores. The answer choice with the highest probability is chosen as the final prediction. For Resolution Prediction, we assign two scores for each answer choice: context+resolution and resolution+outlook, the average of the two scores are used to make comparisons.

Bert For Next Sentence Prediction with Fine-tuned Attention Layers. For each test fold, we finetune the attention layer on the remaining 9 folds of stories and make predictions using a BertForNextSentencePrediction classifier built on top of this pretrained attention layer.

Pretrained Bert For Multiple Choice. We train a BerForMultipleChoice classifier similar to a classifier that solves SWAG (Zellers et al., 2018). For Resolution Prediction full context, we concatenate other stages and outlook with a [SEP] token in between as the input context.

Human Evaluation. We randomly select 125 MCQs for each of the evaluation tasks in full context setting and instruct MTurkers to answer these questions. Only MTurkers who are Masters and obtain a perfect score on our qualification test which consists of five sample MCQs are eligible for the annotation task.

6 Error Analyses

6.1 Distribution of Predictions

Fig. 4 and 5 display the breakdown of predictions for BertForMultipleChoice classifier and average word2vec. Both results justify that conf-climax is the most difficult confounding choice while conf-diff-seed is the easiest with other choices of a difficulty level in between.

6.2 Example of Wrong Prediction

Below is an Outlook Prediction Question that BERT-base MCQ classifier answers incorrectly. This example demonstrates that the ability to predict next sentences is not equivalent to having a social understanding. In this example, all the choices use words that are relevant to the context. The underlying relationship trajectory in each choice, however, varies significantly. A language model is unable to observe the implicit differences among choices and draw the wrong conclusion from only textual information.

Naomi's friend Noah invited her to his house. Naomi was excited to see Noah's new place since the renovations has finally finished. Noah had cooked Naomi's favourite food. Naomi after eating a spoon of it, started to jump here and there in happiness – Naomi loved the dessert! After eating, she expressed her gratitude to Noah for inviting her to the housewarming.

Outlook choices:

correct: *Noah felt grateful for a friend like Naomi. He asked what kind of curtains he should buy in a few weeks.*

conf-climax: *Naomi wanted to get closer to Noah and be his girlfriend.*

conf-buildup: *Naomi would pretend to enjoy the dinner but let quickly stop any of Noah's advances during the evening*

conf-same-seed: *They were both ecstatic that they each wanted a relationship with the other, and they started it that day.*

conf-diff-seed: (Bert's prediction) *Noah likes Naomi very much and wonders if this friendship can develop into something more.*

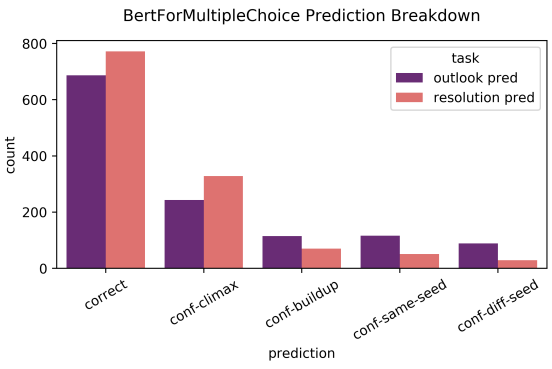

Figure 4: Bert Prediction Distribution.

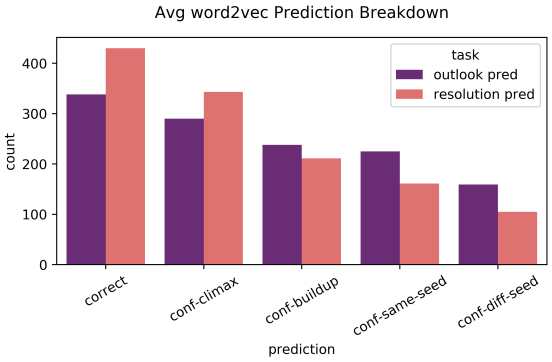

Figure 5: word2vec Prediction Distribution.

	Outlook Prediction		Resolution Prediction	
	Resolution	Full Story	Outlook	Full Story
random	0.20	0.20	0.20	0.20
word2vec	0.3184	0.2736	0.3696	0.3104
BERTNextSent	0.4712	0.3944	0.4832	0.4464
BERTNextSent(finetune)	0.5064	0.4504	0.5096	0.512
BertForMCQ	0.6088	0.592	0.6096	0.6352
Human	N/A	0.80	N/A	0.832

Table 8: Experiment results for baseline models.

6.3 Explanations with Social Elements

We divide the questions into two sets by whether BERT-base MCQ classifier predicts correctly. We compute the difference in VAD (Mohammad, 2018), Affect Intensity (Mohammad, 2017) and SocialSent (Hamilton et al., 2016) scores between each correct and incorrect choice pair and compute the mean for each story, where each choice is represented by the maximum value of lexicon score for individual tokens. We then take the mean of this score across all stories and present the results in Tab. 9, 10 for VAD, Tab. 11, 12 for Affect Intensity and Tab. 13, 14 for SocialSent. For SocialSent, we display the average difference between the correct choice and each confounding choice in the corresponding column. In Tab. 9, we see that incorrectly classified questions have higher arousal difference for outlook prediction whereas for resolution prediction as shown in Tab. 10, incorrectly classified questions have higher valence and dominance difference. This shows that language modeling is unable to extract VAD information from text and use it as a hint in prediction. Similarly, in Tab. 11 and Tab. 12, the greater difference in intensity for joy and anger in outlook prediction incorrect questions and the greater difference in intensity for sadness in resolution prediction incorrect questions imply that mental state is not utilized to make predictions. In Tab. 13 and 14, the greater difference in socially contextualized sentiment score for selective choices in incorrectly answered questions is evidence that sentiment is also not taken into consideration when predictions are made. The results here further prove that language models are unable to pick up underlying social elements including VAD, Affect Intensity and mental states that set the correct answers apart from other choices.

	V	A	D
correct	0.1129	0.1645	0.1325
incorrect	0.1139	0.1742	0.1274

Table 9: Average difference in VAD scores for outlook prediction.

	V	A	D
correct	0.1052	0.1429	0.1168
incorrect	0.1125	0.1463	0.1230

Table 10: Average difference in VAD scores for resolution prediction.

	joy	fear	sadness	anger
correct	0.3182	0.0627	0.1800	0.0653
incorrect	0.3223	0.0660	0.166	0.0726

Table 11: Average difference in affect intensity scores for outlook prediction.

	joy	fear	sadness	anger
correct	0.3562	0.1454	0.2810	0.1420
incorrect	0.3522	0.1492	0.2926	0.1224

Table 12: Average difference in affect intensity scores for resolution prediction.

conf	climax	buildup	same-seed	diff-seed
correct	0.3599	0.3722	0.3985	0.3783
incorrect	0.3778	0.3949	0.3713	0.4026

Table 13: Average difference in SocialSent scores for outlook prediction.

conf	climax	buildup	same-seed	diff-seed
correct	0.3599	0.3722	0.3985	0.3783
incorrect	0.3250	0.3400	0.3725	0.3873

Table 14: Average difference in SocialSent scores in resolution prediction.

7 Summary

We introduce ***Social Narrative Tree***, a corpus of 1250 social narratives that encode rich social elements that collectively drive the relationship trajectory in each story. We propose evaluation tasks Relationship Outlook Prediction MCQ and Resolution Prediction MCQ using this dataset to measure the extent that systems understand social relationships. The tree structure of the dataset provides convenient and meaningful confounding choices, making the tasks challenging for language models. Particularly, the best performance achieved by BERT-base is significantly lower than human performance, justifying that relying on textual information is insufficient in understanding social relationships. A possible future work direction is to incorporate various social elements into predicting models so as to capture the nuances of relationship trajectories in different social situations.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *ArXiv*, abs/1906.05317.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Snigdha Chaturvedi, Dan Goldwasser, and Hal Daume III. 2016. Ask, and shall you receive? understanding desire fulfillment in natural language text. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Steve Duck. 1994. *Meaningful relationships: Talking, sense, and relating*. Sage Publications, Inc.

Gustav Freytag. 1896. *Freytag's technique of the drama: an exposition of dramatic composition and art*. Scholarly Press.

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

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *AAAI*.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. *CoRR*, abs/1606.02820.

Xiaochuang Han, Eunsol Choi, and Chenhao Tan. 2019. No permanent friends or enemies: Tracking relationships between nations from news.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544.

William Labov. 1997. Some further steps in narrative analysis.

Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. 2014. Textblob: simplified text processing. *Secondary TextBlob: simplified text processing*, 3.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif M. Mohammad. 2017. Word affect intensities. *CoRR*, abs/1704.08798.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

G. Prince. 1973. *A Grammar of Stories: An Introduction*. De proprietatibus litterarum. Mouton.

Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang, Pranav Anand, and Marilyn Walker. 2017. Modelling protagonist goals and desires in first-person narrative. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–369.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018a. Modeling naive psychology of characters in simple commonsense stories. *CoRR*, abs/1805.06533.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018b. Event2mind: Commonsense inference on events, intents, and reactions. In *ACL*.

Kenneth Rubin, Cathryn Booth-LaForce, Linda Rose-Krasnor, and Rosemary Mills. 1995. Social relationships and social skills: A conceptual and empirical analysis.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2018. ATOMIC: an atlas of machine commonsense for if-then reasoning. *CoRR*, abs/1811.00146.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *EMNLP*.

R.C. Schank and R. Abelson. 1977. *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Earlbaum Assoc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. *CoRR*, abs/1808.05326.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *CoRR*, abs/1905.07830.