Uncovering Surprising Event Boundaries in Narratives

Anonymous ACL submission

Abstract

When reading stories, people can naturally identify sentences in which a new event starts, i.e., *event boundaries*, using their knowledge of how events typically unfold, but a computational model to detect event boundaries is not yet available. We characterize and detect sentences with expected or surprising event boundaries in an annotated corpus of short diary-like stories, using a model that combines commonsense knowledge and narrative flow features with a RoBERTa classifier. Our results show that, while commonsense and narrative features can help improve performance overall, detecting event boundaries that are more subjective remains challenging for our model. We also find that sentences marking surprising event boundaries are less likely to be causally related to the preceding sentence, but are more likely to express emotional reactions of story characters, compared to sentences with no event boundary.

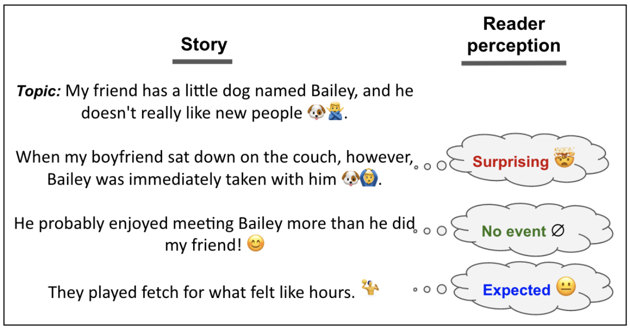

Figure 1: Example story with sentences that contain either a surprising event boundary, no event boundary or an expected event boundary respectively. The annotations of reader perception are from the Hippocorpus dataset (Sap et al., 2022).

1 Introduction

When people read stories, they can easily detect the start of new events through changes in circumstances or in narrative development, i.e., *event boundaries* (Zacks et al., 2007; Bruni et al., 2014; Foster and Keane, 2015; Jafarpour et al., 2019b). These event boundaries can be expected or surprising. For example, in the story in Figure 1 based on crowdsourced annotation, “getting along with a dog who does not generally like new people" marks a *surprising* new event, while “their playing fetch together for a long time" is an *expected* new event.

We aim to study whether machines can detect these surprising or expected event boundaries, using commonsense knowledge and narrative flow features. Characterizing features that are informative in detecting event boundaries can help determine how humans apply expectations on event relationships (Schank and Abelson, 1977; Kurby and Zacks, 2009; Radvansky et al., 2014; Ünal et al., 2019; Zacks, 2020). Furthermore, detection of sentences with event boundaries can also be useful when generating engaging stories with a good amount of surprises. (Yao et al., 2019; Rashkin et al., 2020; Ghazarian et al., 2021).

To differentiate sentences with surprising event boundaries, expected event boundaries, and no event boundaries, we train a classifier using 3925 story sentences with human annotation of event boundaries from diary-like stories about people's everyday lives (Sap et al., 2022). We extract various commonsense and narrative features on relationships between sentences of a story, which can predict the type of event boundaries. Commonsense features include the likelihood that adjacent sentences are linked by commonsense relations from the knowledge graphs Atomic (Sap et al., 2019a) and Glucose (Mostafazadeh et al., 2020). Narrative features include Realis (Sims et al., 2019) that identifies the number of event-related words in a sentence, Sequentiality (Radford et al., 2019; Sap et al., 2022) based on the probability of generating a sentence with varying context and SimGen (Rosset, 2020), which measures the similarity between a sentence and the sentence that is most likely to be generated given the previous sentence. We then combine the prediction based on these features with the prediction from a RoBERTa classifier (Liu et al., 2019), to form overall predictions.

We evaluate the performance of the classification model by measuring F1 of the predictions and compare various configurations of the model to a baseline RoBERTa model. We find that integrating narrative and commonsense features with RoBERTa leads to a significant improvement (+2.2% F1) over a simple RoBERTa classifier. There are also individual differences on the subjective judgment of which sentences contain a surprising or an expected event boundary, that is reflected in the detection model's performance. The performance of our model increases with increasing agreement across the human annotators. Additionally, by interpreting the trained parameters of our model, we find that the absence of causal links between sentences is a strong predictor of surprising event boundaries.

To further analyze how surprising event boundaries relate to deviation from commonsense understanding, we compare the performance of the classification model on the related task of ROC Story Cloze Test (Mostafazadeh et al., 2016). This task concerns whether the ending sentence of a story follows/violates commonsense based on earlier sentences, which can be linked to whether sentences are expected or surprising. Our model performs significantly higher on the ROC Story Cloze Test (87.9% F1 vs 78.0% F1 on our task), showing that surprising event boundaries go beyond merely violating commonsense and therefore can be seen as more challenging to detect. Together, our results suggests that while detecting surprising event boundaries remains a challenging task for machines, a promising direction lies in utilizing commonsense knowledge and narrative features to augment language models.

2 Event Boundary Detection Task

Events have been widely studied in Natural Language Processing. They have often been represented in highly structured formats with word-specific triggers and arguments (Walker et al., 2006; Li et al., 2013; Chen et al., 2017; Sims et al., 2019; Mostafazadeh et al., 2020; Ahmad et al., 2021) or as Subject-Verb-Object-style (SVO) tuples extracted from syntactic parses (Chambers and Jurafsky, 2008; Martin et al., 2018; Rashkin et al., 2018; Sap et al., 2019a). In narratives, events are represented as a continuous flow with multiple boundaries marking new events (Zacks et al., 2007; Graesser et al., 1981; Kurby and Zacks, 2008; Zacks, 2020); however, we lack a model to detect the boundary events that mark the meaningful segmentation of a continuous story into discrete events.

In this work, we study stories from a cognitive angle to detect event boundaries. Such event boundaries relate to our narrative schema understanding (Schank and Abelson, 1977; Chambers and Jurafsky, 2008; Ryan, 2010), commonsense knowledge (Sap et al., 2019a; Mostafazadeh et al., 2020) and world knowledge (Nematzadeh et al., 2018; Bisk et al., 2020). Event boundaries can be surprising or expected based on the knowledge of how a flow of events should unfold. For example, events can be surprising when they deviate from commonsense in terms of what people would predict (e.g., if someone won something, they should not be sad; Sap et al., 2019a). Surprising events can also be low likelihood events (Foster and Keane, 2015) such as seeing someone wear shorts outside in winter, or due to a rapid shift in emotional valence between events (Wilson and Gilbert, 2008) such as seeing a protagonist being defeated. Importantly, there are individual differences in how humans segment narratives into events (Jafarpour et al., 2019a).

We tackle event boundary detection as a three-way classification task that involves distinguishing surprising but plausible event boundaries in story sentences from expected event boundaries and no event boundaries. To mirror how humans read stories, we predict the event boundary label for a sentence using all of its preceding sentences in the story, as well as the general story topic as context. *Surprising* event boundaries are novel events that are unexpected given their context, such as a dog getting along with someone despite not typically liking new people. *Expected* event boundaries are novel events that are not surprising, such as a person playing a new game with a dog for a long time given that they like each other. In contrast, sentences with *no event* boundary typically continue or elaborate on the preceding event, such as a person liking a dog given that they get along with the dog (Figure 1).

3 Event-annotated Data

We use the event-annotated sentences from stories in the Hippocorpus dataset to study event boundaries. This dataset contains 240 diary-like sto-

Majority label	#Samples (%)	% majority agreement (std)
No event	2255 (57.5)	68.1 (13.9)
Expected	650 (16.6)	58.8 (10.6)
Surprising	509 (13.0)	61.7 (11.9)
Tied	511 (13.0)	41.1 (5.7)
Total	3925 (100)	62.2 (15.2)

Table 1: Descriptive Statistics for Event-Annotated sentences. Majority label refers to the most common annotation of a sample from 8 independent annotators. If there is a tie between 2 labels, it is categorized as tied. Majority agreement is the proportion of sample annotations for the majority label.

ries about everyday life experiences, which annotated at the sentence level (Sap et al., 2022). Stories were inspected for the absence of offensive or person-identifying content. For the annotation, eight crowdworkers were shown a story sentence by sentence and were asked to mark whether each sentence contained a new surprising or expected event boundary, or no event boundary at all, based on their subjective judgment (Sap et al., 2022). Summarized in Table 1, based on the majoritarian vote, most sentences (57.5%) contain no event boundaries while 16.6% and 13.0% of sentences contains expected and surprising event boundaries, respectively.

Due to the inherent subjectivity of the task, aggregating labels into a majority label yields low agreement (e.g., 61.7% for surprising event boundaries; Table 1). Therefore, at training time, we use the proportion of annotations for each event boundary type as the label instead of the majority vote, because such distributional information is a better reflection of the inherent disagreement among human judgements (Pavlick and Kwiatkowski, 2019). At test time, we use the majority vote as a gold label, since measuring performance on distribution modelling is less intuitive to interpret, and subsequently break down performance by agreement level to take disagreements into account.

4 Event Boundary Detection Model

We first describe informative commonsense and narrative features that we extract for the event boundary detection model. Then, we describe how we integrate these features with a RoBERTa classifier in our model before detailing our experimental setup. Figure 2 depicts an overview of our model.

4.1 Features

We select a collection of commonsense features (Atomic and Glucose relations) and narrative flow features (Realis, Sequentiality and SimGen). A model is trained separately from our main model for Atomic relations, Glucose relations and Realis while models for Sequentiality and SimGen are used without further training. Features of story sentences are extracted as input into the main model. Because language modelling alone might not be sufficient to learn such features (Gordon and Van Durme, 2013; Sap et al., 2019a), we provide the extracted features to the model instead of relying on the language models to learn them implicitly.

Atomic relations are event relations from a social commonsense knowledge graph containing numerous events that can be related to one another (Sap et al., 2019a). The event relations in this graph consists of:

Emotional **React**ion,
The **Effect** of an event,
Want to do after the event,
What **Need**s to be done before an event,
The **Intent**ion to do a certain event,
What **Attr**ibutes an event expresses.

When an event affects the subject, the feature name is preceded by an `x`, while if it affects others, it has an `o`. For example, an `xWant` of a sentence *PersonX pays PersonY a compliment* is that *PersonX will want to chat with PersonY*, and an `oWant` is that *PersonY will compliment PersonX back*. We use Atomic relations because surprising event boundaries can involve breaches of commonsense understanding (Bosselut et al., 2019; Sap et al., 2019a; Mostafazadeh et al., 2020; Gabriel et al., 2021). Furthermore, some Atomic relations (`xReact` and `oReact`) concern emotional affect and therefore can be used to capture changes in emotional valence, which can cause events to be seen as surprising (Wilson and Gilbert, 2008).

We train an Atomic relation classifier using a RoBERTa-base model (Liu et al., 2019) to classify event-pairs into one of the nine possible relationship labels as well as a None label (to introduce negative samples). We achieved a validation F1 of 77.15%, which is high for a 10-way classification task. We describe training and other experimental details in the Appendix. When making inferences on the event-annotated dataset, we predict the likelihood that a preceding sentence in a story will be

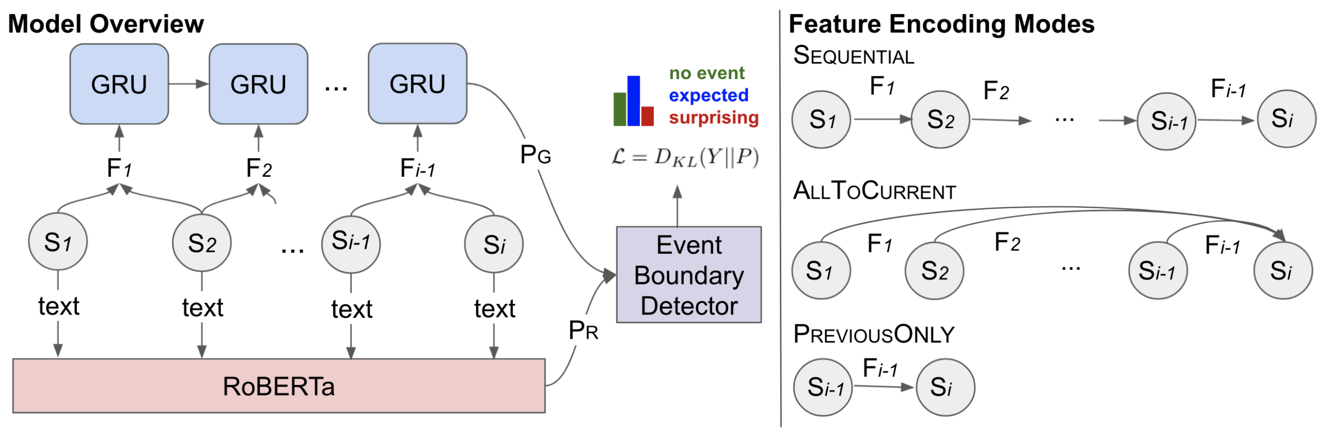

Figure 2: (Left) Our model involves a **GRU** to combine features from sentence pairs with three feature encoding modes, **RoBERTa** to consider story sentences and **Event Boundary Detector** to combine predictions made by the two components. S_n and F_n refer to sentence n and features n respectively, while P_G and P_R are predictions made by the GRU and RoBERTa. The output is a probability distribution over no event boundary, expected event boundary and surprising event boundary, which is used to update model parameters together with the label using the **Kullback-Leibler Divergence** loss function. (Right) **Features** (Atomic, Glucose, Realis, Sequentiality and SimGen) can be extracted as input into the GRU in three feature encoding modes: SEQUENTIAL (shown in Model Overview), ALLTOCURRENT and PREVIOUSONLY.

related to the current sentence via each of the nine relationship label. Because Atomic relations are directed relations (e.g., *I ate some cake* `xEffect` *I am full* is different from *I am full* `xEffect` *I ate some cake*), we also made the reverse inference in case commonsense relations between sentences exist in the reverse direction. Together, 9 forward atomic relation features and 9 reverse features (marked with'-r') are used.

Glucose relations are event relations from another commonsense knowledge dataset containing relations between event-pairs in 10 dimensions (Mostafazadeh et al., 2020). Glucose relation features are used to complement Atomic relation features in its coverage of commonsense relations. Dim-1 to 5 are described below while Dim-6 to 10 are the reverse/passive form of Dim-1 to 5 respectively.

`Dim-1`: **Event** that causes/enables
`Dim-2`: **Emotion**/human drive that motivates
`Dim-3`: **Change in location** that enables
`Dim-4`: **State of possession** that enables
`Dim-5`: **Other attribute** that enables

Glucose relation classifier was trained on a RoBERTa-base model to classify event-pairs from its annotated dataset into one of ten possible relation labels as well as a None label. We used the *specific* version of Glucose events represented in natural language. As a result, we achieved a validation F1 of 80.94%. Training and other experimental details are in the Appendix. During inference on the Event-annotated dataset, we predict and use as features the likelihood that the current sentence will be related to a preceding sentence via each relation label.

Realis events are words that serve as triggers (i.e., head words) for structured event representations (Sims et al., 2019). Realis event words denote concrete events that actually happened, meaning that a higher number of Realis event words suggests greater likelihood of the sentence containing a new event boundary (expected or surprising). We trained a BERT-base model (Devlin et al., 2019) on an annotated corpus of literary novel extracts (Sims et al., 2019). We achieved a validation F1 of 81.85%, inspired by and on par with Sap et al. (2020). Then, we use the trained model to make inference on story sentences in the Event-annotated dataset. Finally, we used the number of `Realis` words in each sentence as a feature. Training and other experimental details are in the Appendix.

Sequentiality is a measure of the difference in conditional negative log-likelihood of generating a sentence given the previous sentence or otherwise (Sap et al., 2020, 2022). Sequentiality can be a predictor for unlikely events, which can cause surprise (Foster and Keane, 2015). We use GPT-2 (Radford et al., 2019) to measure this negative log-likelihood since it is a Left-to-Right model, which matches the order in which annotators were shown sentences in a story. NLL of each sentence was obtained in two different contexts. `NLL_topic` is based on the sentence alone with only the topic as prior context, while `NLL_topic+prev` uses the previous sentence as additional context to study the link between adjacent sentences. Finally, `Sequentiality` is obtained by taking their difference. Experimental details are in the Appendix.

$$NLL_{topic} = -\frac{1}{|s_i|} \log p_{LM}(s_i \mid Topic)$$

$$NLL_{topic+prev} = -\frac{1}{|s_i|} \log p_{LM}(s_i \mid Topic, s_{i-1})$$

SimGen is computed as the cosine similarity between each sentence and the most likely generated sentence given the previous sentence, under a large Left-to-Right language model (specifically, Turing-NLG; Rosset, 2020). Then, we separately converted the original sentence and generated sentence into sentence embeddings using a pre-trained MPnet-base model (Song et al., 2020). Finally, the generated embeddings and the original embeddings are compared for cosine similarity, which is used as a feature. Experimental details are in the Appendix.

4.2 Model Architecture

We propose a model to integrate feature-based prediction with language-based prediction of event boundaries, illustrated in Figure 2 (left). The predictions are independently made with extracted features using a gated recurrent unit (GRU) and with language (i.e., story sentences) using RoBERTa. Then these predictions are combined into a final predicted distribution for the three types of event boundaries. Our model is then trained using the Kullback-Leibler Divergence loss.

GRU is used to combine features relating the current sentence i to prior sentences in a story. It sequentially considers information concerning prior sentences, which mimics the annotator's procedure of identifying event boundaries as they read one sentence at the time. As seen in Figure 2 (right), we use three feature encoding modes to determine the features that are used as input into the GRU, as inspired by literature on event segmentation (Pettijohn and Radvansky, 2016; Baldassano et al., 2018; Zacks, 2020). These three modes represents different ways of facilitating information flow between sentences, which can have distinct effects on identifying event boundaries.

The first mode, SEQUENTIAL, encodes features from all previous sentences in the story in a recurrent way (1 to 2, 2 to 3 ... $i-1$ to i) up until the current sentence i. The second mode, ALLTOCURRENT, uses features from each of the previous sentences to the current sentence i (1 to i, 2 to i ... $i-1$ to i). The third mode, PREVIOUSONLY, ($i-1$ to i) only feeds into the GRU the features relating to the previous sentence. For all modes, the dimension of each time step input is K_G, representing the total number of distinct features. We then project the final output of the GRU, $h_G \in \mathbb{R}^{K_G}$, into a 3-dimensional vector space representing the unnormalized probability distribution over event boundary types.

RoBERTa is used to make predictions based on text in story sentences. We use all story sentences up to sentence i inclusive. We then project the hidden state of the first token, $h_R \in \mathbb{R}^{K_R}$, into a 3-dimensional space representing the unnormalized probability distribution over event boundary types.

Combining predictions We combine predictions made by the GRU (P_G) and RoBERTa (P_R) by concatenating their predictions and multiplying it with a linear classifier of size (6, 3) to output logits of size (3). The logits are then normalized using Softmax to give a distribution of the three types of event boundaries (P). The weights of the linear classifier are initialized by concatenating two identity matrix of size 3 ($\mathbf{I}_3$), which serves to perform elementwise addition between the predictions of the GRU and RoBERTa at early stages of the training process.

$$W := [\mathbf{I}_3; \mathbf{I}_3] \quad (1)$$

$$P := Softmax(W([P_G; P_R])) \quad (2)$$

Loss function We use the Kullback-Leibler Divergence loss function to train the model. We use it over the standard Cross Entropy loss function because our training targets are in the form: proportion of annotations for each type of event boundary (e.g., 0.75, 0.125, 0.125 for no event, expected and surprising respectively). Including such distributional information in our training targets over using the majority annotation only can reflect the inherent disagreement among human judgements (Pavlick and Kwiatkowski, 2019), which is important to capture for event boundaries given that they are subjective judgements.

4.3 Experimental setup

We seek to predict the event-boundary annotation for each Hippocorpus story sentence, using preceding sentences in the story as context, as shown in Figure 2. Additional training and experimental details are available in the Appendix.

K-fold Cross-validation Because of the limited size of the dataset (n=3925), we split the dataset in k-folds (k=10), using one fold (n=392) for validation and nine other folds combined for training. From each of the 10 models, we obtained the prediction for the validation set. Together, the validation sets for the 10 models combine to form predictions for the entire dataset, which we use to conduct significance testing in order to compare the performance of models.

GRU was accessed from PyTorch, with K_G set to 33 and a hidden dimension of 33.

RoBERTa RoBERTa-base-uncased was used, accessed from HuggingFace Transformers library (with 12-layer, 768-hidden (K_R), 12-heads, 110M parameters, 0.1 dropout). When more than 10 prior sentences are available in a story, we use only the most recent 10 sentences due to RoBERTa input sequence length limitations.

Evaluation Metrics While capturing distributional information of subjective judgement labels (Pavlick and Kwiatkowski, 2019) is important for training, it can also be difficult to interpret for evaluation. Therefore, we decided to predict for the most likely label during evaluation and compare it against the majority label for each sample. Some samples do not have a single majority label (e.g., equal number of expected and surprising annotations) and these samples were excluded. We use micro-averaged F1 as the metric.

5 Results and Discussion

We first quantify the performance of our model in detecting event boundaries, using a coarse-grained performance measure on F1 with respect to majority vote. Then, we investigate how the performance varies based on annotation subjectivity. Finally, we inspect the model parameters to identify commonsense and narrative features that are most informative in detecting event boundaries.

Improving prediction of event boundaries As seen in Table 2, RoBERTa alone performs fairly well in predicting event boundaries (F1 = 75.8%, within 2.2% F1 of our best performing model), but can be further supported by our commonsense and narrative features to improve its performance. In contrast, the commonsense and narrative features alone do not perform as well.[1] Overall, our best performing set-up is the Event Detector (PREVIOUSONLY) with F1 = 78.0%, which is significantly different from RoBERTa alone based on McNemar's test ($p < 0.05$). [2] Its overall strong performance is largely contributed by its strong performance in detecting no event boundaries and expected event boundaries. F1 for no event boundary is higher than for both surprising and expected event boundaries, likely because there are more sentences with no event boundaries as seen in Table 1. The PREVIOUSONLY configuration performs best for no event boundaries and expected event boundaries likely because determining whether the current sentence continues an expected event (or not) requires retaining the latest information in working memory (Jafarpour et al., 2019a). However, the SEQUENTIAL configuration seems to perform the best in predicting surprising event boundaries. Compared to no/expected event boundaries, we hypothesize that predicting surprising event boundaries requires taking into account how the story developed prior to the previous sentence in setting up the context for the current sentence. This

	overall F1	no event F1	expected F1	surprising F1
Event Detector (w RoBERTa)				
- PREVIOUSONLY*	78.0	87.2	60.0	59.7
- SEQUENTIAL	77.3	86.6	57.5	60.5
- ALLTOCURRENT	76.9	86.3	57.5	59.7
RoBERTa	75.8	86.2	55.8	54.3
Event Detector (w/o RoBERTa)				
- ALLTOCURRENT	63.9	81.8	32.3	24.8
- SEQUENTIAL	63.8	82.1	34.6	19.5
- PREVIOUSONLY	63.4	81.8	31.8	21.2

Table 2: Event detection task: Performance of Event Detector compared to baseline model. *: significant difference from RoBERTa based on McNemar's test ($p < 0.05$)

[1] We also increased learning rate to 1e-3 for better performance given the absence of RoBERTa predictions in this ablation set-up

[2] McNemar's test is used to determine whether samples that have been predicted accurately (or not) by one model overlap with those that have predicted accurately (or not) by another model

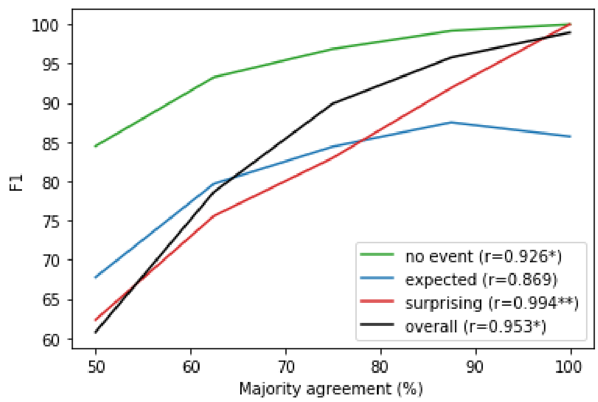

Figure 3: F1 by Event Detector (PREVIOUSONLY) against majority agreement, on all 10 folds. * means that Pearson's r is significant at $p < 0.05$ and ** at $p < 0.001$.

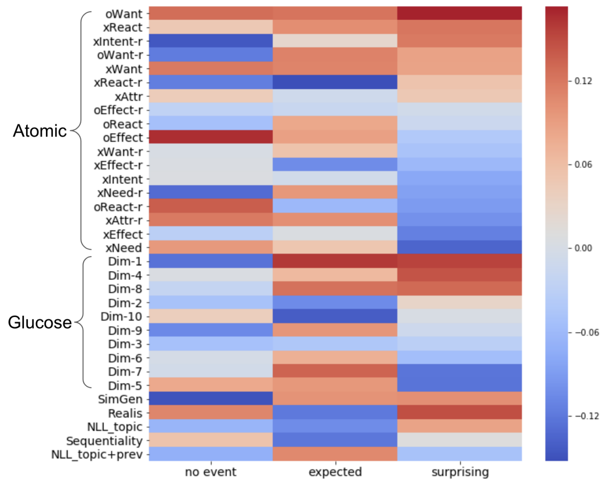

Figure 4: Feature weights towards each label in GRU component of Event Detector (PREVIOUSONLY)

finding echoes results by Townsend (2018) that showed that surprising sentences take long time to read because it requires changing our mental model formed from previous sentences.

F1 varies with majority agreement Since the annotations were subjective and did not always agree, we further examine our best model's performance (PREVIOUSONLY) with respect to annotation agreement. As shown in Figure 3, F1 increases with majority label agreement (Pearson's $r = 0.953$, $p < 0.05$). Such positive correlations are observed across all event boundary labels (Pearson's r = 0.869-0.994) and is especially strong for surprising event boundaries (Pearson's $r = 0.994$, $p < 0.001$). This means that most errors are made on samples that have low agreement among annotators. For example to show this contrast, after "*She and I are very close so it was great to see her marrying someone she loves*," 7 out of 8 annotators indicated that "*The most memorable moment was when I spilled champagne on my dress before the wedding*" was surprising. On the other hand, after "*It was a hot day in July that our community decided to paint a mural on an intersection for public art*," only 4 out of 8 annotators indicated that "*I had decided to volunteer to help paint.*" was surprising. The results suggest that our model performance reflects the variability and agreements in humans annotations of event boundaries. We hypothesize that the event boundaries with more agreement are based on features that are shared across the annotators, such as commonsense knowledge; therefore, the model performs well in detecting those. Whereas, our model struggles with detecting event boundaries that are more subjective.

Predictive features By integrating a separate feature-based classifier, the Event Boudary Detector model allows us to examine the model parameters and determine features that are associated with surprising, expected or no event boundaries. First, we take the average of the GRU classifier weights for each of the 10 cross-validated models. Then, we plot these weights for each label in Figure 4, and summarize the findings below.

Features that relate to commonsense relations: `oEffect`, `xEffect` and `Glucose Dim-6` (caused by) are most predictive of expected event boundaries. This can indicate that events that are an effect of/caused by a prior event can be expected by annotators, as also noted by Graesser et al. (1981). An example of an expected event boundary is *"I told her we could go for coffee sometime."*, as an effect of *"We had a good time together."* `xNeed` is least indicative of surprising event boundaries. This is likely because `xNeed` refers to what the subject need to do before an activity, which is procedural and unlikely to cause surprise. An example is *"I was grocery shopping a few weeks ago."* which is needed before *"I had purchased my items and was leaving the store."*

Features that explain unlikely events `Realis` is highest for surprising event boundaries, suggesting that surprising event boundaries tend to contain the most concrete event-words. Surprising event boundaries also have the highest likelihood when conditioned on the story topic (`NLL_topic`) while expected events are highest when conditioned based on the topic and the previous sentence

(NLL_topic+prev). This suggests that surprising events are often inline with the story topic but not with the previous sentence. Therefore, the low likelihood of transitioning between the previous and current sentence is a strong predictor of surprising event boundaries, in line with findings by Foster and Keane (2015) on how the difficulty of linking two adjacent events is an important factor in causing surprise.

Features that explain changes in emotional valence Compared to sentences that contain no event boundaries, sentences that contain either expected or surprising event boundaries have higher xReact and oReact, which are emotional responses either by the subject or by others to an event. For example, this is the case for the surprising and emotional event boundary "*I remember it was like the 3rd or 4th game when something bad happened.*." This suggests that event boundaries are more likely when a sentence is more emotionally charged, echoing work by Dunsmoor et al. (2018) on how event segmentation is particularly frequent when the emotion of fear is triggered.

6 Comparison with Story Cloze Test

To better understand how surprising event boundaries relate to deviation from commonsense reasoning, we compare our Event Boundary Detection Task to the ROC Story Cloze Test (Mostafazadeh et al., 2016). This Test involves identifying whether a candidate ending sentence follows commonsense (*commonsense ending*) or deviates from commonsense (*nonsense ending*) given the first four sentences of a short story. Deviation from commonsense reasoning is one factor that can cause surprise (Sap et al., 2019a) and therefore comparing our task to the ROC Story Cloze Test can allow us to potentially isolate deviations from commonsense from other factors that can cause surprise. The ROC Story Cloze Test dataset contains 3142 samples with 1571 commonsense endings and 1571 nonsense endings.[3] We train a separate Event Boundary Detector model on the ROC Story Cloze Test, using the same experimental setup as for event boundary detection, except the loss function; we use the cross-entropy loss since only one label is available for each sample.[4]

[3]We use the Winter 2018 version, which contains a dev and a test set. As in previous work (Schwartz et al., 2017), we train our model on the dev portion.

[4]Training takes 20 minutes on an Nvidia P100 GPU.

	overall F1	nonsense ending F1	commonsense ending F1
Event Detector w RoBERTa			
- AllToCurrent	87.9	87.8	88.0
- PreviousOnly	87.6	87.3	87.8
- Sequential	87.3	87.1	87.5
RoBERTa	87.7	87.6	87.8

Table 3: ROC Story Cloze Test

Performance of Event Detector on ROC Story Cloze Test Compared to the Event Boundary Detection task, models perform significantly better on the ROC Story Cloze Test (highest F1 = 78.0% vs. 87.9%, $p < 0.001$ based on a two-tailed t-test, as observed in Tables 2 and 3). While the tasks are not directly comparable due to the inherent subjectivity of the Event Boundary Detection Task, the higher performance on the ROC Story Cloze Test suggests that identifying surprising, expected or no event boundaries may be more challenging than identifying commonsense or nonsense endings. Our commonsense and narrative features also do not seem to significantly improve upon RoBERTa's performance in the ROC Story Cloze Test (+0.2% F1). This indicates that detecting whether a story ending follows commonsense can be effectively approached using RoBERTa alone, making it relatively easier to tackle.

7 Conclusion

We tackle the task of identifying event boundaries in stories. We propose a model that combines predictions made using commonsense and narrative features with a RoBERTa classifier. We found that integrating commonsense and narrative features can significantly improve the prediction of surprising event boundaries through detecting violations to commonsense relations (especially relating to the absence of causality), low likelihood events, and changes in emotional valence. Our model is capable in detecting event boundaries with high annotator agreement but limited in detecting those with lower agreement. Compared to identifying commonsense and nonsense story endings in Story Cloze Test, our task is found to be more challenging. Our results suggest that considering commonsense knowledge and narrative features can be a promising direction towards characterizing and detecting event boundaries in stories.

References

Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Gate: Graph attention transformer encoder for cross-lingual relation and event extraction.

Christopher Baldassano, Uri Hasson, and Kenneth A. Norman. 2018. Representation of real-world event schemas during narrative perception. *The Journal of Neuroscience*, 38(45):9689–9699.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Luis Emilio Bruni, Sarune Baceviciute, and Mohammed Arief. 2014. Narrative cognition in interactive systems: Suspense-surprise and the p300 erp component. In *Interactive Storytelling*, pages 164–175, Cham. Springer International Publishing.

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.

Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically labeled data generation for large scale event extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 409–419, Vancouver, Canada. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph E. Dunsmoor, Marijn C. W. Kroes, Caroline M. Moscatelli, Michael D. Evans, Lila Davachi, and Elizabeth A. Phelps. 2018. Event segmentation protects emotional memories from competing experiences encoded close in time. *Nature Human Behaviour*, 2(4):291–299.

Meadhbh I. Foster and Mark T. Keane. 2015. Predicting surprise judgments from explanation graphs.

Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz, Ronan Le Bras, Maxwell Forbes, and Yejin Choi. 2021. Paragraph-level commonsense transformers with recurrent memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12857–12865.

Sarik Ghazarian, Zixi Liu, Akash S M, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2021. Plot-guided adversarial example construction for evaluating open-domain story generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4334–4344, Online. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, page 25–30, New York, NY, USA. Association for Computing Machinery.

Arthur C Graesser, Scott P Robertson, and Patricia A Anderson. 1981. Incorporating inferences in narrative representations: A study of how and why. *Cognitive Psychology*, 13(1):1–26.

Anna Jafarpour, Elizabeth A Buffalo, Robert T Knight, and Anne GE Collins. 2019a. Event segmentation reveals working memory forgetting rate. *Available at SSRN 3614120*.

Anna Jafarpour, Sandon Griffin, Jack J Lin, and Robert T Knight. 2019b. Medial orbitofrontal cortex, dorsolateral prefrontal cortex, and hippocampus differentially represent the event saliency. *Journal of cognitive neuroscience*, 31(6):874–884.

CA Kurby and JM Zacks. 2009. Segmentation in the perception and memory of events. *Trends in cognitive sciences*.

Christopher A Kurby and Jeffrey M Zacks. 2008. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 12(2):72–79.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, Sofia, Bulgaria. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Lara Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark Riedl. 2018. Event representations for automated story generation with deep neural nets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

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.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Kyle A. Pettijohn and Gabriel A. Radvansky. 2016. Narrative event boundaries, reading times, and expectation. *Memory & Cognition*, 44(7):1064–1075.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Gabriel A. Radvansky, Andrea K. Tamplin, Joseph Armendarez, and Alexis N. Thompson. 2014. Different kinds of causality in event cognition. *Discourse Processes*, 51(7):601–618.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295, Online. Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Corby Rosset. 2020. Turing-nlg: A 17-billion-parameter language model by microsoft.

Marie-Laure Ryan. 2010. Narratology and cognitive science: A problematic relation. *Style*, 44(4):469–495.

Maarten Sap, Eric Horvitz, Yejin Choi, Noah A Smith, and James W Pennebaker. 2020. Recollection versus imagination: Exploring human memory and cognition via neural language models. In *ACL*.

Maarten Sap, Anna Jafarpour, Yejin Choi, Noah A. Smith, James W. Pennebaker, and Eric Horvitz. 2022. Computational lens on cognition: Study of autobiographical versus imagined stories with large-scale language models.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

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

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. In *CoNLL*.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.

David J. Townsend. 2018. Stage salience and situational likelihood in the formation of situation models during sentence comprehension. *Lingua*, 206:1–20.

Ercenur Ünal, Yue Ji, and Anna Papafragou. 2019. From event representation to linguistic meaning. *Topics in Cognitive Science*, 13(1):224–242.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus.

Timothy D. Wilson and Daniel T. Gilbert. 2008. Explaining away: A model of affective adaptation. *Perspectives on Psychological Science*, 3(5):370–386. PMID: 26158955.

Lili Yao, Nanyun Peng, Weischedel Ralph, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*.

Jeffrey M. Zacks. 2020. Event perception and memory. *Annual Review of Psychology*, 71(1):165–191. PMID: 31905113.

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. 2007. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273.

A Appendix

A.1 Atomic relations training details

We used the train/dev/test splits from the original Atomic dataset. Negative samples are created by matching a Atomic event node to a corresponding tail event node from another sample based on the relationship involved. Sepcifically, negative sampling was performed on groups (['xWant', 'oWant', 'xNeed', 'xIntent'],['xReact', 'oReact', 'xAttr'],['xEffect', 'oEffect']) given that the tail event nodes in each group are more similar, creating more discriminating negative samples, as inspired by Sap et al. (2019b). One negative sample is introduced every nine positive samples, since there are nine labels. We used a learning rate of 1e-4, batch size of 64, 8 epochs and AdamW optimizer. Training took 18 hours on a Nvidia P100 GPU.

A.2 Glucose relations training details

Because the Glucose dataset was not split initially, we randomly split the dataset into train/dev/test splits based on a 80/10/10 ratio. For each sample in Glucose, annotations share similar head event nodes in Dim-1 to 5 and similar tail event nodes in Dim-6 to 10. Therefore, our negative sampling strategy for Dim-1 to 5 involves randomly choosing a tail node from Dim-6 to 10 and vice-versa. As a result, one negative sample is introduced every five samples. During training, we used a learning rate of 1e-4, batch size of 64, 8 epochs and AdamW optimizer. Training took 15 hours on a Nvidia P100 GPU.

A.3 Realis training details

We used the train/dev/test split from the Realis dataset. During training, we used the AdamW optimizer, a learning rate of 2e-5, 3 epochs and batch size of 4, as inspired by (Sap et al., 2020). Training took 1 hour on a Nvidia P100 GPU.

A.4 Sequentiality experimental details

GPT2-small was accessed from HuggingFace Transformers library and used without further fine-tuning. It has 125M parameters, a context window of 1024, hidden state dimension of 768, 12 heads and dropout of 0.1.

A.5 SimGen experimental details

We used the Turing-NLG model without further fine-tuning. The model has 17B and we used it with top-p sampling (top-p=0.85), temperature=1.0 and max sequence length of 64 tokens. MPnet-base model was accessed from the Sentence-BERT library (Reimers and Gurevych, 2019) and used without further fine-tuning.

A.6 Event Boundary Detection Model training details

AdamW optimizer was used with $\alpha = 5{*}10^{-6}$, following a uniform search using F1 as the criterion at intervals of $\{2.5, 5, 7.5, 10\} * 10^n; -6 \leq n \leq -3$. Learning rate was linearly decayed (8 epochs) with 100 warm-up steps. Batch size of 16 was used. Validation was done every 0.25 epochs during training. Training each model took around 30 minutes on an Nvidia P100 GPU.