Modelling Suspense in Short Stories as Uncertainty Reduction over Neural Representation

David Wilmot and **Frank Keller**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
david.wilmot@ed.ac.uk, keller@inf.ed.ac.uk

Abstract

Suspense is a crucial ingredient of narrative fiction, engaging readers and making stories compelling. While there is a vast theoretical literature on suspense, it is computationally not well understood. We compare two ways for modelling suspense: surprise, a backward-looking measure of how unexpected the current state is given the story so far; and uncertainty reduction, a forward-looking measure of how unexpected the continuation of the story is. Both can be computed either directly over story representations or over their probability distributions. We propose a hierarchical language model that encodes stories and computes surprise and uncertainty reduction. Evaluating against short stories annotated with human suspense judgements, we find that uncertainty reduction over representations is the best predictor, resulting in near human accuracy. We also show that uncertainty reduction can be used to predict suspenseful events in movie synopses.

1 Introduction

As current NLP research expands to include longer, fictional texts, it becomes increasingly important to understand narrative structure. Previous work has analyzed narratives at the level of characters and plot events (e.g., Gorinski and Lapata, 2018; Martin et al., 2018). However, systems that process or generate narrative texts also have to take into account what makes stories compelling and enjoyable. We follow a literary tradition that makes *And then?* (Forster, 1985; Rabkin, 1973) the primary question and regards suspense as a crucial factor of storytelling. Studies show that suspense is important for keeping readers' attention (Khrypko and Andreae, 2011), promotes readers' immersion and suspension of disbelief (Hsu et al., 2014), and plays a big part in making stories enjoyable and interesting (Oliver, 1993; Schraw et al., 2001). Computationally less well understood, suspense has only sporadically been used in story generation systems (O'Neill and Riedl, 2014; Cheong and Young, 2014).

Suspense, intuitively, is a feeling of anticipation that something risky or dangerous will occur; this includes the idea both of uncertainty and jeopardy. Take the play *Romeo and Juliet:* Dramatic suspense is created throughout — the initial duel, the meeting at the masquerade ball, the marriage, the fight in which Tybalt is killed, and the sleeping potions leading to the death of Romeo and Juliet. At each moment, the audience is invested in something being at stake and wonders how it will end.

This paper aims to model suspense in computational terms, with the ultimate goal of making it deployable in NLP systems that analyze or generate narrative fiction. We start from the assumption that concepts developed in psycholinguistics to model human language processing at the word level (Hale, 2001, 2006) can be generalised to the story level to capture suspense, the Hale model. This assumption is

similar concepts to model suspense in games (Ely et al., 2015; Li et al., 2018), the Ely model. Common to both approaches is the idea that suspense is a form of expectation: In games, we expect to win or lose instead in stories, we expect that the narrative will end a certain way.

We will therefore compare two ways for modelling narrative suspense: surprise, a backward-looking measure of how unexpected the current state is given the story so far; and uncertainty reduction, a forward-looking and measure of how unexpected the continuation of the story is. Both measures can be computed either directly over story representations, or indirectly over the probability distributions over such representations. We propose a hierarchical language model based on Generative Pre-Training (GPT, Radford et al., 2018) to encode story-level representations and develop an

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1763–1788
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

inference scheme that uses these representations to compute both surprise and uncertainty reduction. For evaluation, we use the WritingPrompt corpus of short stories (Fan et al., 2018), part of which we annotate with human sentence-by-sentence judgements of suspense. We find that surprise over representations and over probability distributions both predict suspense judgements. However uncertainty reduction over representations is better, resulting in near human-level accuracy. We also show that our models can be used to predict turning points, i.e., major narrative events, in movie synopses (Papalampidi et al., 2019).

2 Related Work

In narratology, uncertainty over outcomes is traditionally seen as suspenseful (e.g., O'Neill, 2013; Zillmann, 1996; Abbott, 2008). Other authors claim that suspense can exist without uncertainty (e.g., Smuts, 2008; Hoeken and van Vliet, 2000; Gerrig, 1989) and that readers feel suspense even when they read a story for the second time (Delatorre et al., 2018), which is unexpected if suspense is uncertainty; this is referred to as the paradox of suspense (Prieto-Pablos, 1998; Yanal, 1996). Considering *Romeo and Juliet* again, in the first view suspense is motivated by primarily by uncertainty over what will happen. Who will be hurt or killed in the fight? What will happen after marriage? However, at the beginning of the play we are told "from forth the fatal loins of these two foes, a pair of star-crossed lovers take their life", and so the suspense is more about being invested in the plot than not knowing the outcome, aligning more with the second view: suspense can exist without uncertainty. We do not address the paradox of suspense directly in this paper, but we are guided by the debate to operationalise methods that encompass both views. The Hale model is closer to the traditional model of suspense as being about uncertainty. In contrast, the Ely model is more in line with the second view that uncertainty matters less than consequentially different outcomes.

In NLP, suspense is studied most directly in natural language generation, with systems such as Dramatis (O'Neill and Riedl, 2014) and Suspenser (Cheong and Young, 2014), two planning-based story generators that use the theory of Gerrig and Bernardo (1994) that suspense is created when a protagonist faces obstacles that reduce successful outcomes. Our approach, in contrast, models suspense using general language models fine-tuned on stories, without planning and domain knowledge. The advantage is that the model can be trained on large volumes of available narrative text without requiring expensive annotations, making it more generalisable.

Other work emphasises the role of characters and their development in story understanding (Bamman et al., 2014, 2013; Chaturvedi et al., 2017; Iyyer et al., 2016) or summarisation (Gorinski and Lapata, 2018). A further important element of narrative structure is plot, i.e., the sequence of events in which characters interact. Neural models have explicitly modelled events (Martin et al., 2018; Harrison et al., 2017; Rashkin et al., 2018) or the results of actions (Roemmele and Gordon, 2018; Liu et al., 2018a,b). On the other hand, some neural generation models (Fan et al., 2018) just use a hierarchical model on top of a language model; our architecture follows this approach.

3 Models of Suspense

3.1 Definitions

In order to formalise measures of suspense, we assume that a story consists of a sequence of sentences. These sentences are processed one by one, and the sentence at the current timepoint t is represented by an embedding e_t (see Section 4 for how embeddings are computed). Each embedding is associated with a probability $P(e_t)$. Continuations of the story are represented by a set of possible next sentences, whose embeddings are denoted by e^i_{t+1}.

The first measure of suspense we consider is surprise (Hale, 2001), which in the psycholinguistic literature has been successfully used to predict word-based processing effort (Demberg and Keller, 2008; Roark et al., 2009; Van Schijndel and Linzen, 2018a,b). Surprise is a backward-looking predictor: it measures how unexpected the current word is given the words that preceded it (i.e., the left context). Hale formalises surprise as the negative log of the conditional probability of the current word. For stories, we compute surprise over sentences. As our sentence embeddings e_t include information about the left context $e_1, \ldots, e_{t-1}$, we can write *Hale surprise* as:

$$S_t^{\text{Hale}} = -\log P(e_t) \tag{1}$$

An alternative measure for predicting word-by-word processing effort used in psycholinguistics is entropy reduction (Hale, 2006). This measure is forward-looking: it captures how much the current word changes our expectations about the words we will encounter next (i.e., the right context). Again, we compute entropy at the story level, i.e., over sentences instead of over words. Given a probability distribution over possible next sentences $P(e^i_{t+1})$, we calculate the entropy of that distribution. Entropy reduction is the change of that entropy from one sentence to the next:

$$\begin{aligned} H_t &= -\sum_i P(e^i_{t+1}) \log P(e^i_{t+1}) \\ U^{\text{Hale}}_t &= H_{t-1} - H_t \end{aligned} \tag{2}$$

Note that we follow Frank (2013) in computing entropy over surface strings, rather than over parse states as in Hale's original formulation.

In the economics literature, Ely et al. (2015) have proposed two measures that are closely related to Hale surprise and entropy reduction. At the heart of their theory of suspense is the notion of belief in an end state. Games are a good example: the state of a tennis game changes with each point being played, making a win more or less likely. Ely et al. define surprise as the amount of change from the previous time step to the current time step. Intuitively, large state changes (e.g., one player suddenly comes close to winning) are more surprising than small ones. Representing the state at time t as e_t, *Ely surprise* is defined as:

$$S^{\text{Ely}}_t = (e_t - e_{t-1})^2 \tag{3}$$

Ely et al.'s approach can be adapted for modelling suspense in stories if we assume that each sentence in a story changes the state (the characters, places, events in a story, etc.). States e_t then become sentence embeddings, rather than beliefs in end states, and Ely surprise is the distance between the current embedding e_t and the previous embedding e_{t-1}. In this paper, we will use L1 and L2 distances; other authors (Li et al., 2018) experiment with information gain and KL divergence, but found worse performance when modelling suspense in games. Just like Hale surprise, Ely surprise models backward-looking prediction, but over representations, rather than over probabilities.

Ely et al. also introduce a measure of forward-looking prediction, which they define as the expected difference between the current state e_t and the next state e_{t+1}:

$$\begin{aligned} U^{\text{Ely}}_t &= \mathbb{E}[(e_t - e^i_{t+1})^2] \\ &= \sum_i P(e^i_{t+1})(e_t - e^i_{t+1})^2 \end{aligned} \tag{4}$$

This is closely related to Hale entropy reduction, but again the entropy is computed over states (sentence embeddings in our case), rather than over probability distributions. Intuitively, this measure captures how much the uncertainty about the rest of the story is reduced by the current sentence. We refer to the forward-looking measures in Equations (2) and (4) as *Hale and Ely uncertainty reduction,* respectively.

Ely et al. also suggest versions of their measures in which each state is weighted by a value α_t, thus accounting for the fact that some states may be more inherently suspenseful than others:

$$\begin{aligned} S^{\alpha\text{Ely}}_t &= \alpha_t(e_t - e_{t-1})^2 \\ U^{\alpha\text{Ely}}_t &= \mathbb{E}[\alpha_{t+1}(e_t - e^i_{t+1})^2] \end{aligned} \tag{5}$$

We stipulate that sentences with high emotional valence are more suspenseful, as emotional involvement heightens readers' experience of suspense. This can be captured in Ely et al.'s framework by assigning the αs the scores of a sentiment classifier.

3.2 Modelling Approach

We now need to show how to compute the surprise and uncertainty reduction measures introduced in the previous section. This involves building a model that processes stories sentence by sentence, and assigns each sentence an embedding that encodes the sentence and its preceding context, as well as a probability. These outputs can then be used to compute a surprise value for the sentence.

Furthermore, the model needs to be able to generate a set of possible next sentences (story continuations), each with an embedding and a probability. Generating upcoming sentences is potentially very computationally expensive since the number of continuations grows exponentially with the number of future time steps. As an alternative, we can therefore sample possible next sentences from a corpus and use the model to assign them embeddings and probabilities. Both of these approaches will produce sets of upcoming sentences, which we can then use to compute uncertainty reduction. While we have so far only talked about the next sentences, we will also experiment with uncertainty reduction computed using longer rollouts.

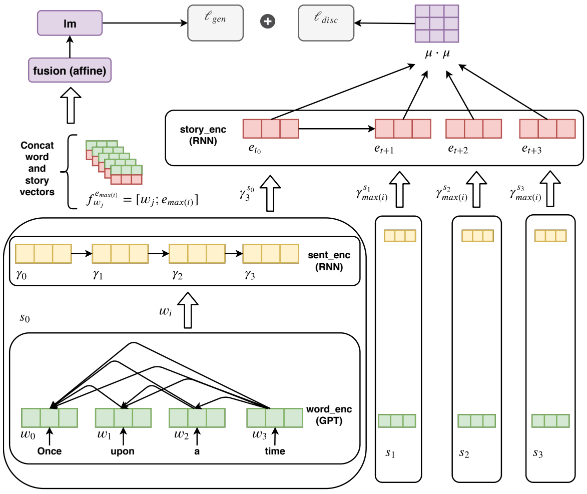

Figure 1: Architecture of our hierarchical model. See text for explanation of the components *word_enc*, *sent_enc*, and *story_enc*.

4 Model

4.1 Architecture

Our overall approach leverages contextualised language models, which are a powerful tool in NLP when pretrained on large amounts of text and fine tuned on a specific task (Peters et al., 2018; Devlin et al., 2019). Specifically, we use Generative Pre-Training (GPT, Radford et al., 2018), a model which has proved successful in generation tasks (Radford et al., 2019; See et al., 2019).

Hierarchical Model Previous work found that hierarchical models show strong performance in story generation (Fan et al., 2018) and understanding tasks (Cai et al., 2017). The language model and hierarchical encoders we use are unidirectional, which matches the incremental way in which human readers process stories when they experience suspense.

Figure 1 depicts the architecture of our hierarchical model.[1] It builds a chain of representations that anticipates what will come next in a story, allowing us to infer measures of suspense. For a given sentence, we use GPT as our word encoder (*word_enc* in Figure 1) which turns each word in a sentence into a word embedding w_i. Then, we use an RNN (*sent_enc*) to turn the word embeddings of the sentences into a sentence embedding γ_i. Each sentence is represented by the hidden state of its last word, which is then fed into a second RNN (*story_enc*) that computes a story embedding. The overall story representation is the hidden state of its last sentence. Crucially, this model also gives us e_t, a contextualised representation of the current sentence at point t in the story, to compute surprise and uncertainty reduction.

Model training includes a generative loss ℓ_{gen} to improve the quality of the sentences generated by the model. We concatenate the word representations w_j for all word embeddings in the latest sentence with the latest story embedding $e_{\max(t)}$. This is run through affine ELU layers to produce enriched word embedding representations, analogous to the Deep Fusion model (Gülçehre et al., 2015), with story state instead of a translation model. The related Cold Fusion approach (Sriram et al., 2018) proved inferior.

Loss Functions To obtain the discriminatory loss ℓ_{disc} for a particular sentence s in a batch, we compute the dot product of all the story embeddings e in the batch, and then take the cross-entropy across the batch with the correct next sentence:

$$\ell_{disc}(e_{t+1}^{i=s}) = -\log \frac{\exp(e_{t+1}^{i=s} \cdot e_t)}{\sum_i \exp(e_{t+1}^i \cdot e_t)} \qquad (6)$$

Modelled on Quick Thoughts (Logeswaran and Lee, 2018), this forces the model to maximise the dot product of the correct next sentence versus other sentences in the same story, and negative examples from other stories, and so encourages representations that anticipate what happens next.

The generative loss in Equation (7) is a standard LM loss, where w_j is the GPT word embeddings from the sentence and $e_{\max(t)}$ is the story context that each word is concatenated with:

$$\ell_{gen} = -\sum_j \log P(w_j | w_{j-1}, w_{j-2}, \ldots; e_{\max(t)}) \quad (7)$$

The overall loss is $\ell_{disc} + \ell_{gen}$. More advanced generation losses (e.g., Zellers et al., 2019) could be used, but are an order of magnitude slower.

4.2 Inference

We compute the measures of surprise and uncertainty reduction introduced in Section 3.1 using the output of the story encoder *story_enc*. In addition to the contextualised sentence embeddings e_t, this requires their probabilities $P(e_t)$, and a distribution over alternative continuations $P(e_{t+1}^i)$.

We implement a recursive beam search over a tree of future sentences in the story, looking between one and three sentences ahead (rollout). The

[1] Model code and scripts for evaluation are available at https://github.com/dwlmt/Story-Untangling/tree/acl-2020-dec-submission

probability is calculated using the same method as the discriminatory loss, but with the cosine similarity rather than the dot product of the embeddings e_t and e^i_{t+1} fed into a softmax function. We found that cosine outperformed dot product on inference as the resulting probability distribution over continuations is less concentrated.

5 Methods

Dataset The overall goal of this work is to test whether the psycholinguistic and economic theories introduced in Section 3 are able to capture human intuition of suspense. For this, it is important to use actual stories which were written by authors with the aim of being engaging and interesting. Some of the story datasets used in NLP do not meet this criterion; for example ROC Cloze (Mostafazadeh et al., 2016) is not suitable because the stories are very short (five sentences), lack naturalness, and are written by crowdworkers to fulfill narrow objectives, rather than to elicit reader engagement and interest. A number of authors have also pointed out technical issues with such artificial corpora (Cai et al., 2017; Sharma et al., 2018).

Instead, we use WritingPrompts (Fan et al., 2018), a corpus of circa 300k short stories from the `/r/WritingPrompts` subreddit. These stories were created as an exercise in creative writing, resulting in stories that are interesting, natural, and of suitable length. The original split of the data into 90% train, 5% development, and 5% test was used. Pre-processing steps are described in Appendix A.

Annotation To evaluate the predictions of our model, we selected 100 stories each from the development and test sets of the WritingPrompts corpus, such that each story was between 25 and 75 sentence in length. Each sentence of these stories was judged for narrative suspense; five master workers from Amazon Mechanical Turk annotated each story after reading instructions and completing a training phase. They read one sentence at a time and provided a suspense judgement using the five-point scale consisting of Big Decrease in suspense (1% of the cases), Decrease (11%), Same (50%), Increase (31%), and Big Increase (7%). In contrast to prior work (Delatorre et al., 2018), a relative rather than absolute scale was used. Relative judgements are easier to make while reading, though in practice, the suspense curves generated are very similar, with a long upward trajectory and flattening or dip near the end. After finishing a story, annotators had to write a short summary of the story.

	GRU	LSTM
Loss	5.84	5.90
Discriminatory Acc.	0.55	0.54
Discriminatory Acc. $k = 10$	0.68	0.68
Generative Acc.	0.37	0.46
Generative Acc. $k = 10$	0.85	0.85
Cosine Similarity	0.48	0.50
L2 Distance	1.73	1.59
Number of Epochs	4	2

Table 1: For accuracy the baseline probability is 1 in 99; $k = 10$ is the accuracy of the top 10 sentences of the batch. From the best epoch of training on the WritingPrompts development set.

In the instructions, suspense was framed as *dramatic tension*, as pilot annotations showed that the term suspense was too closely associated with murder mystery and related genres. Annotators were asked to take the character's perspective when reading to achieve stronger inter-annotator agreement and align closely with literary notions of suspense. During training, all workers had to annotate a test story and achieve 85% accuracy before they could continue. Full instructions and the training story are in Appendix B.

The inter-annotator agreement α (Krippendorff, 2011) was 0.52 and 0.57 for the development and test sets, respectively. Given the inherently subjective nature of the task, this is substantial agreement. This was achieved after screening out and replacing annotators who had low agreement for the stories they annotated (mean $\alpha < 0.35$), showed suspiciously low reading times (mean RT < 600 ms per sentence), or whose story summaries indicated low-quality annotation.

Training and Inference The training used SGD with Nesterov momentum (Sutskever et al., 2013) with a learning rate of 0.01 and a momentum of 0.9. Models were run with early stopping based on the mean of the accuracies of training tasks. For each batch, 50 sentence blocks from two different stories were chosen to ensure that the negative examples in the discriminatory loss include easy (other stories) and difficult (same story) sentences.

We used the pretrained GPT weights but fine-tuned the encoder and decoder weights on our task. For the RNN components of our hierarchical model, we experimented with both GRU (Chung et al.,

2015) and LSTM (Hochreiter and Schmidhuber, 1997) variants. The GRU model had two layers in both *sen_enc* and *story_enc*; the LSTM model had four layers each in *sen_enc* and *story_enc*. Both had two fusion layers and the size of the hidden layers for both model variants was 768. We give the results of both variants on the tasks of sentence generation and sentence discrimination in Table 1. Both perform similarly, with slightly worse loss for the LSTM variant, but faster training and better generation accuracy. Overall, model performance is strong: the LSTM variant picks out the correct sentence 54% of the time and generates it 46% of the time. This indicates that our architecture successfully captures the structure of stories.

At inference time, we obtained a set of story continuations either by random sampling or by generation. Random sampling means that n sentences were selected from the corpus and used as continuations. For generation, sentences were generated using top-k sampling (with $k = 50$) using the GPT language model and the approach of Radford et al. (2019), which generates better output than beam search (Holtzman et al., 2018) and can outperform a decoder (See et al., 2019). For generation, we used up to 300 words as context, enriched with the story sentence embeddings from the corresponding points in the story. For rollouts of one sentence, we generated 100 possibilities at each step; for rollouts of two, 50 possibilities and rollouts of three, 25 possibilities. This keeps what is an expensive inference process manageable.

Importance We follow Ely et al. in evaluating weighted versions of their surprise and uncertainty reduction measure $S_t^{\alpha\text{Ely}}$ and $U_t^{\alpha\text{Ely}}$ (see Equation (5)). We obtain the α_t values by taking the sentiment scores assigned by the VADER sentiment classifier (Hutto and Gilbert, 2014) to each sentence and multiplying them by 1.0 for positive sentiment and 2.0 for negative sentiment. The stronger negative weighting reflects the observation that negative consequences can be more important than positive ones (O'Neill, 2013; Kahneman and Tversky, 2013).

Baselines We test a number of baselines as alternatives to surprise and uncertainty reduction derived from our hierarchical model. These baselines also reflect how much change occurs from one sentence to the next in a story: WordOverlap is the Jaccard similarity between the two sentences, GloveSim is the cosine similarity between the averaged Glove (Pennington et al., 2014) word embeddings of the two sentences, and GPTSim is the cosine similarity between the GPT embeddings of the two sentences. The α baseline is the weighted VADER sentiment score.

6 Results

6.1 Narrative Suspense

Task The annotator judgements are relative (amount of decrease/increase in suspense from sentence to sentence), but the model predictions are absolute values. We could convert the model predictions into discrete categories, but this would fail to capture the overall arc of the story. Instead, we convert the relative judgements into absolute suspense values, where $J_t = j_1 + \cdots + j_t$ is the absolute value for sentence t and $j_1, \ldots, j_t$ are the relative judgements for sentences 1 to t. We use -0.2 for Big Decrease, -0.1 for Decrease, 0 for Same, 0.1 for Increase, and 0.2 for Big Increase.[2] Both the absolute suspense judgements and the model predictions are normalised by converting them to z-scores.

To compare model predictions and absolute suspense values, we use Spearman's ρ (Sen, 1968) and Kendall's τ (Kendall, 1975). Rank correlation is preferred because we are interested in whether human annotators and models view the same part of the story as more or less suspenseful; also, rank correlation methods are good at detecting trends. We compute ρ and τ between the model predictions and the judgements of each of the annotators (i.e., five times for five annotators), and then take the average. We then average these values again over the 100 stories in the test or development sets. As the human upper bound, we compute the mean pairwise correlation of the five annotators.

Results Figure 2 shows surprise and uncertainty reduction measures and human suspense judgements for an example story (text and further examples in Appendix C). We performed model selection using the correlations on the development set, which are given in Table 2. We experimented with all the measures introduced in Section 3.1, computing sets of alternative sentences either us-

[2] These values were fitted with predictions (or cross-worker annotation) using 5-fold cross validation and an L1 loss to optimise the mapping. A constraint is placed so that Same is 0, increases are positive and decreases are negative with a minimum 0.05 distance between.

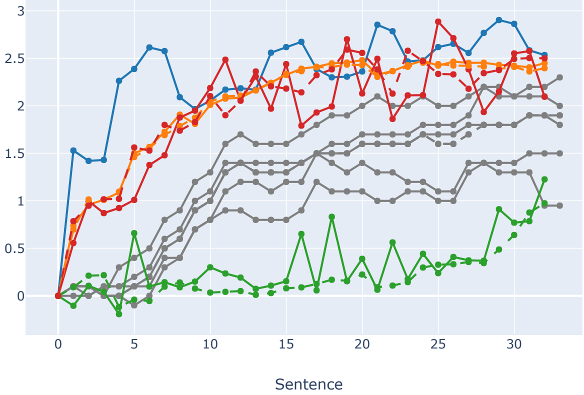

Figure 2: Story 27, **Human**, S^{Hale}, S^{Ely}, U^{Ely}, $U^{\alpha\text{Ely}}$. Solid lines: generated alternative continuations, dashed lines: sampled alternative continuations.

ing generated continuations (Gen) or continuations sampled from the corpus (Cor), except for S^{Ely}, which can be computed without alternatives. We compared the LSTM and GRU variants (see Section 4) and experimented with rollouts of up to three sentences. We tried L1 and L2 distance for the Ely measures, but only report L1, which always performed better.

Discussion On the development set (see Table 2), we observe that all baselines perform poorly, indicating that distance between simple sentence representations or raw sentiment values do not model suspense. We find that Hale surprise S^{Hale} performs well, reaching a maximum ρ of .675 on the development set. Hale uncertainty reduction U^{Hale}, however, performs consistently poorly. Ely surprise S^{Ely} also performs well, reaching as similar value as Hale surprise. Overall, Ely uncertainty reduction U^{Ely} is the strongest performer, with ρ = .698, numerically outperforming the human upper bound.

Some other trends are clear from the development set: using GRUs reduces performance in all cases but one; rollout of more than one never leads to an improvement; sentiment weighting (prefix α in the table) always reduces performance, as it introduces considerable noise (see Figure 2). We therefore eliminate the models that correspond to these settings when we evaluate on the test set.

For the test set results in Table 3 we also report upper and lower confidence bounds computed using the Fisher Z-transformation ($p < 0.05$). On the test set, U^{Ely} again is the best measure, with a correlation statistically indistinguishable from human performance (based on CIs). We find that absolute correlations are higher on the test set, presumably reflecting the higher human upper bound.

Prediction	**Model**	**Roll**	τ ↑	ρ ↑
Human			**.553**	**.614**
Baselines	WordOverlap	1	.017	.026
	GloveSim	1	.017	.029
	GPTSim	1	.021	.031
	α	1	**.024**	**.036**
S^{Hale}-Gen	GRU	1	.145	.182
	LSTM	1	**.434**	**.529**
S^{Hale}-Cor	GRU	1	.177	.214
	LSTM	1	**.580**	**.675**
U^{Hale}-Gen	GRU	1	**.036**	**.055**
	LSTM	1	.009	.016
U^{Hale}-Cor	GRU	1	.048	.050
	LSTM	1	**.066**	**.094**
S^{Ely}	GRU	1	**.484**	**.607**
	LSTM	1	.427	.539
$S^{\alpha\text{Ely}}$	GRU	1	.089	.123
	LSTM	1	**.115**	**.156**
U^{Ely}-Gen	GRU	1	.241	.161
		2	.304	.399
	LSTM	1	**.610**	**.698**
		2	.393	.494
U^{Ely}-Cor	GRU	1	.229	.264
		2	.512	.625
		3	.515	.606
	LSTM	1	**.594**	**.678**
		2	.564	.651
		3	.555	.645
$U^{\alpha\text{Ely}}$-Gen	GRU	1	.216	.124
		2	.219	.216
	LSTM	1	**.474**	**.604**
		2	.316	.418
$U^{\alpha\text{Ely}}$-Cor	GRU	1	.205	.254
		2	.365	.470
	LSTM	1	**.535**	**.642**
		2	.425	.534

Table 2: Development set results for WritingPrompts for generated (Gen) or corpus sampled (Cor) alternative continuations; α indicates sentiment weighting. **Bold:** best model in a given category; **red:** best model overall.

Overall, we conclude that our hierarchical architecture successfully models human suspense

Prediction	$\tau \uparrow$	$\rho \uparrow$
Human	**.652** (.039)	**.711** (.033)
S^{Hale}-Gen	.407 (.089)	.495 (.081)
S^{Hale}-Cor	**.454** (.085)	**.523** (.079)
U^{Hale}-Gen	.036 (.102)	.051 (.102)
U^{Hale}-Cor	.061 (.100)	.088 (.101)
S^{Ely}	.391 (.092)	.504 (.082)
U^{Ely}-Gen	**.620** (.067)	**.710** (.053)
U^{Ely}-Cor	.605 (.069)	.693 (.056)

Table 3: Test set results for WritingPrompts for generated (Gen) or corpus sampled (Cor) continuations. LSTM with rollout one; brackets: confidence intervals.

judgements on the WritingPrompts dataset. The overall best predictor is U^{Ely}, uncertainty reduction computed over story representations. This measure combines the probability of continuation (S^{Hale}) with distance between story embeddings (S^{Ely}), which are both good predictors in their own right. This finding supports the theoretical claim that suspense is an expectation over the change in future states of a game or a story, as advanced by Ely et al. (2015).

6.2 Movie Turning Points

Task and Dataset An interesting question is whether the peaks in suspense in a story correspond to important narrative events. Such events are sometimes called turning points (TPs) and occur at certain positions in a movie according to screenwriting theory (Cutting, 2016). A corpus of movie synopses annotated with turning points is available in the form of the TRIPOD dataset (Papalampidi et al., 2019). We can therefore test if surprise or uncertainty reduction predict TPs in TRIPOD. As our model is trained on a corpus of short stories, this will also serve as an out-of-domain evaluation.

Papalampidi et al. (2019) assume five TPs: 1. Opportunity, 2. Change of Plans, 3. Point of no Return, 4. Major Setback, and 5. Climax. They derive a prior distribution of TP positions from their test set, and use this to constrain predicted turning points to windows around these prior positions. We follow this approach and select as the predicted TP the sentence with the highest surprise or uncertainty reduction value within a given constrained window. We report the same baselines as in the previous experiment, as well as the Theory Baseline, which uses screenwriting theory to predict where in a movie a given TP should occur (e.g., Point of No Return theoretically occurs 50% through the movie). This baseline is hard to beat (Papalampidi et al., 2019).

	Dev $D \downarrow$	**Test** $D \downarrow$
Human	Not reported	**4.30 (3.43)**
Theory Baseline	9.65 (0.94)	7.47 (3.42)
TAM	**7.11 (1.71)**	6.80 (2.63)
WordOverlap	13.9 (1.45)	12.7 (3.13)
GloveSim	**10.2 (0.74)**	**10.4 (2.54)**
GPTSim	16.8 (1.47)	18.1 (4.71)
α	11.3 (1.24)	11.2 (2.67)
S^{Hale}-Gen	**8.27 (0.68)**	**8.72 (2.27)**
U^{Hale}-Gen	10.9 (1.02)	10.69 (3.66)
S^{Ely}	9.54 (0.56)	9.01 (1.92)
$S^{\alpha\text{Ely}}$	9.95 (0.78)	9.54 (2.76)
U^{Ely}-Gen	8.75 (0.76)	8.38 (1.53)
U^{Ely}-Cor	8.74 (0.76)	8.50 (1.69)
$U^{\alpha\text{Ely}}$-Gen	8.80 (0.61)	7.84 (3.34)
$U^{\alpha\text{Ely}}$-Cor	**8.61 (0.68)**	**7.78 (1.61)**

Table 4: TP prediction on the TRIPOD development and test sets. D is the normalised distance to the gold standard; CI in brackets.

Results and Discussion Figure 3 plots both gold standard and predicted TPs for a sample movie synopsis (text and further examples in Appendix D). The results on the TRIPOD development and test sets are reported in Table 4 (we report both due to the small number of synopses in TRIPOD). We use our best LSTM model with a of rollout of one; the distance measure for Ely surprise and uncertainty reduction is now L2 distance, as it outperformed L1 on TRIPOD. We report results in terms of D, the normalised distance between gold standard and predicted TP positions.

On the test set, the best performing model with $D = 7.78$ is $U^{\alpha\text{Ely}}$-Cor, with $U^{\alpha\text{Ely}}$-Gen only slightly worse. It is outperformed by TAM, the best model of Papalampidi et al. (2019), which however requires TP annotation at training time. $U^{\alpha\text{Ely}}$-Cor is close to the Theory Baseline on the test set, an impressive result given that our model has no TP supervision and is trained on a different domain. The fact that models with sentiment

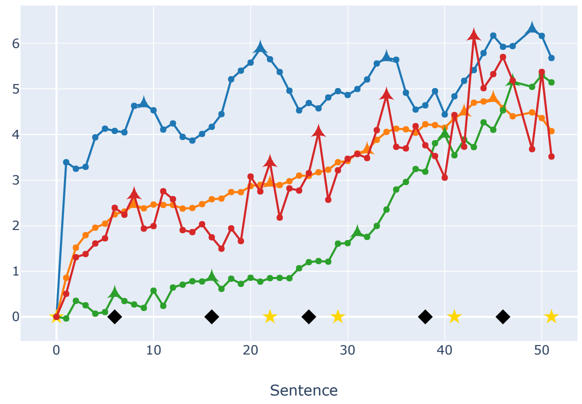

Figure 3: Movie 15 Minutes, S^{Hale}, S^{Ely}, U^{Ely}, $U^{\alpha\text{Ely}}$, ◆ theory baseline, ★ TP annotations, triangles are predicted TPs.

weighting (prefix α) perform well here indicates that turning points often have an emotional resonance as well as being suspenseful.

7 Conclusions

Our overall findings suggest that by implementing concepts from psycholinguistic and economic theory, we can predict human judgements of suspense in storytelling. That uncertainty reduction (U^{Ely}) outperforms probability-only (S^{Hale}) and state-only (S^{Ely}) surprise suggests that, while consequential state change is of primary importance for suspense, the probability distribution over the states is also a necessary factor. Uncertainty reduction therefore captures the view of suspense as reducing paths to a desired outcome, with more consequential shifts as the story progresses (O'Neill and Riedl, 2014; Ely et al., 2015; Perreault, 2018). This is more in line with the Smuts (2008) Desire-Frustration view of suspense, where uncertainty is secondary.

Strong psycholinguistic claims about suspense are difficult to make due to several weaknesses in our approach, which highlight directions for future research: the proposed model does not have a higher-level understanding of event structure; most likely it picks up the textual cues that accompany dramatic changes in the text. One strand of further work is therefore analysis: Text could be artificially manipulated using structural changes, for example by switching the order of sentences, mixing multiple stories, including a summary at the beginning that foreshadows the work, masking key suspenseful words, or paraphrasing. An analogue of this would be adversarial examples used in computer vision. Additional annotations, such as how certain readers are about the outcome of the story, may also be helpful in better understanding the relationship between suspense and uncertainty. Automated interpretability methods as proposed by Sundararajan et al. (2017), could shed further light on models' predictions.

The recent success of language models in wide-ranging NLP tasks (e.g., Radford et al., 2019) has shown that language models are capable of learning semantically rich information implicitly. However, generating plausible future continuations is an essential part of the model. In text generation, Fan et al. (2019) have found that explicitly incorporating coreference and structured event representations into generation produces more coherent generated text. A more sophisticated model would incorporate similar ideas.

Autoregressive models that generate step by step alternatives for future continuations are computationally impractical for longer rollouts and are not cognitively plausible. They also differ from the Ely et al. (2015) conception of suspense, which is in terms of Bayesian beliefs over a longer-term future state, not step by step. There is much recent work (e.g., Ha and Schmidhuber (2018); Gregor et al. (2019)), on state-space approaches that model beliefs as latent states using variational methods. In principle, these would avoid the brute-force calculation of a rollout and conceptually, anticipating longer-term states aligns with theories of suspense.

Related tasks such as inverting the understanding of suspense to utilise the models in generating more suspenseful stories may also prove fruitful.

This paper is a baseline that demonstrates how modern neural network models can implicitly represent text meaning and be useful in a narrative context without recourse to supervision. It provides a springboard to further interesting applications and research on suspense in storytelling.

Acknowledgments

The authors would like to thank the anonymous reviewers, Pinelopi Papalampidi and David Hodges for reviews of the annotation task, the AMT annotators, and Mirella Lapata, Ida Szubert, and Elizabeth Nielson for comments on the paper. Wilmot's work is funded by an EPSRC doctoral training award.

References

H Porter Abbott. 2008. *The Cambridge introduction to narrative*. Cambridge University Press.

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

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending: Strong neural baselines for the roc story cloze task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 616–622.

Snigdha Chaturvedi, Mohit Iyyer, and Hal Daume III. 2017. Unsupervised learning of evolving relationships between literary characters. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Yun-Gyung Cheong and R Michael Young. 2014. Suspenser: A story generation system for suspense. *IEEE Transactions on Computational Intelligence and AI in Games*, 7(1):39–52.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *International Conference on Machine Learning*, pages 2067–2075.

James E Cutting. 2016. Narrative theory and the dynamics of popular movies. *Psychonomic Bulletin & Review*, 23(6):1713–1743.

Pablo Delatorre, Carlos León, Alberto G Salguero, Manuel Palomo-Duarte, and Pablo Gervás. 2018. Confronting a paradox: a new perspective of the impact of uncertainty in suspense. *Frontiers in Psychology*, 9:1392.

Vera Demberg and Frank Keller. 2008. Data from eyetracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 101(2):193–210.

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.

Jeffrey Ely, Alexander Frankel, and Emir Kamenica. 2015. Suspense and surprise. *Journal of Political Economy*, 123(1):215–260.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *ACL*.

Edward Morgan Forster. 1985. *Aspects of the Novel*, volume 19. Houghton Mifflin Harcourt.

Stefan L Frank. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, 5(3):475–494.

Richard J Gerrig. 1989. Suspense in the absence of uncertainty. *Journal of Memory and Language*, 28(6):633–648.

Richard J Gerrig and Allan BI Bernardo. 1994. Readers as problem-solvers in the experience of suspense. *Poetics*, 22(6):459–472.

Philip John Gorinski and Mirella Lapata. 2018. What's this movie about? a joint neural network architecture for movie content analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1770–1781, New Orleans, Louisiana. Association for Computational Linguistics.

Karol Gregor, George Papamakarios, Frederic Besse, Lars Buesing, and Theophane Weber. 2019. Temporal difference variational auto-encoder. In *International Conference on Learning Representations*.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

David Ha and Jürgen Schmidhuber. 2018. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems 31*, pages 2451–2463. Curran Associates, Inc. `https://worldmodels.github.io`.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, volume 2, pages 159–166, Pittsburgh, PA. Association for Computational Linguistics.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive science*, 30(4):643–672.

Brent Harrison, Christopher Purdy, and Mark O Riedl. 2017. Toward automated story generation with markov chain monte carlo methods and deep neural networks. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hans Hoeken and Mario van Vliet. 2000. Suspense, curiosity, and surprise: How discourse structure influences the affective and cognitive processing of a story. *Poetics*, 27(4):277–286.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649, Melbourne, Australia. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

C. T. Hsu, M. Conrad, and A. M. Jacobs. 2014. Fiction feelings in harry potter: haemodynamic response in the mid-cingulate cortex correlates with immersive reading experience. *NeuroReport*, 25:1356–1361.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

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.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific.

MG Kendall. 1975. Rank correlation measures. *Charles Griffin, London*, 202:15.

Y. Khrypko and P. Andreae. 2011. Towards the problem of maintaining suspense in interactive narrative. In *Proceedings of the 7th Australasian Conference on Interactive Entertainment*, pages 5:1–5:3.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. 2011. *Annenberg School for Communication Departmental Papers: Philadelphia*.

Zhiwei Li, Neil Bramley, and Todd M. Gureckis. 2018. Modeling dynamics of suspense and surprise. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018, Madison, WI, USA, July 25-28, 2018*.

Chunhua Liu, Haiou Zhang, Shan Jiang, and Dong Yu. 2018a. DEMN: Distilled-exposition enhanced matching network for story comprehension. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Fei Liu, Trevor Cohn, and Timothy Baldwin. 2018b. Narrative modeling with memory chains and semantic supervision. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 278–284, Melbourne, Australia. Association for Computational Linguistics.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Lara J Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O Riedl. 2018. Event representations for automated story generation with deep neural nets. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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.

M. B. Oliver. 1993. Exploring the paradox of the enjoyment of sad films. *Human Communication Research*, 19:315–342.

Brian O'Neill. 2013. *A computational model of suspense for the augmentation of intelligent story generation*. Ph.D. thesis, Georgia Institute of Technology.

Brian O'Neill and Mark Riedl. 2014. Dramatis: A computational model of suspense. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 944–950.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie plot analysis via turning point identification. 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 1707–1717, Hong Kong, China. Association for Computational Linguistics.

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

Joseph Perreault. 2018. *The Universal Structure of Plot Content: Suspense, Magnetic Plot Elements, and the Evolution of an Interesting Story*. Ph.D. thesis, University of Idaho.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Juan A Prieto-Pablos. 1998. The paradox of suspense. *Poetics*, 26(2):99–113.

Eric S Rabkin. 1973. *Narrative suspense." When Slim turned sideways..."*. Ann Arbor: University of Michigan Press.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Https://openai.com/blog/language-unsupervised/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Singapore. Association for Computational Linguistics.

Melissa Roemmele and Andrew Gordon. 2018. An encoder-decoder approach to predicting causal relations in stories. In *Proceedings of the First Workshop on Storytelling*, pages 50–59, New Orleans, Louisiana. Association for Computational Linguistics.

G. Schraw, Flowerday, T., and S. Lehman. 2001. Increasing situational interest in the classroom. *Educational Psychology Review*, 13:211–224.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019. Do massively pretrained language models make better storytellers? *arXiv preprint arXiv:1909.10705*.

Pranab Kumar Sen. 1968. Estimates of the regression coefficient based on kendall's tau. *Journal of the American statistical association*, 63(324):1379–1389.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the story ending biases in the story cloze test. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 752–757.

Aaron Smuts. 2008. The desire-frustration theory of suspense. *The Journal of Aesthetics and Art Criticism*, 66(3):281–290.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2018. Cold fusion: Training seq2seq models together with language models. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, pages 387–391.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147.

Marten Van Schijndel and Tal Linzen. 2018a. Can entropy explain successor surprisal effects in reading? *CoRR*, abs/1810.11481.

Marten Van Schijndel and Tal Linzen. 2018b. Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018, Madison, WI, USA, July 25-28, 2018*.

Robert J Yanal. 1996. The paradox of suspense. *The British Journal of Aesthetics*, 36(2):146–159.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *CoRR*, abs/1905.12616.

Dolf Zillmann. 1996. The psychology of suspense in dramatic exposition. *Suspense: Conceptualizations, theoretical analyses, and empirical explorations*, pages 199–231.

A Pre-processing

WritingPrompts comes from a public forum of short stories and so is naturally noisy. Story authors often use punctuation in unusual ways to mark out sentences or paragraph boundaries and there are lots of spelling mistakes. Some of these cause problems with the GPT model and in some circumstances can cause it to crash. To improve the quality, sentence demarcations are left as they are from the original WritingPrompts dataset but some sentences are cleaned up and others skipped over. Skipping over is also why there sometimes are gaps in the graph plots as the sentence was ignored during training and inference. The pre-processing steps are as follows. Where substitutions are made rather than ignoring the sentence, the token is replaced by the Spacy (Honnibal and Montani, 2017) POS tag.

1. **English Language:** Some phrases in sentences can be non-English, Whatthelang (Joulin et al., 2016) is used to filter out these sentences.

2. **Nondictionary words:** *PyDictionary* and *PyEnchant* and used to check if each word is a dictionary word. If not they are replaced.

3. **Repeating Symbols:** Some author mark out sections by using a string of characters such as *************** or !!!!!!!!!!!!. This can cause the Pytorch GPT implementation to break so repeating characters are replaced with a single one.

4. **Ignoring sentences:** If after all of these replacements there are not three or more GPT word pieces ignoring the POS replacements then the sentence is skipped. The same processing applies to generating sentences in the inference. Occasionally the generated sentences can be nonsense, so the same criteria are used to exclude them.

B Mechanical Turk Written Instructions

These are the actual instructions given to the Mechanical Turk Annotators, plus the example in Table 5:

INSTRUCTIONS For the first HIT there will be an additional training step to pass. This will take about 5 minutes. After this you will receive a code which you can enter in the code box to bypass the training for subsequent HITS. Other stories are in separate HITS, please search for ”Story dramatic tension, reading sentence by sentence” to find them. The training completion code will work for all related HITS.

You will read a short story and for each sentence be asked to assess how the dramatic tension increases, decreases or stays the same. Each story will take an estimated 8-10 minutes. Judge each sentence on how the dramatic tension has changed as felt by the main characters in the story, not what you as a reader feel. Dramatic tension is the excitement or anxiousness over what will happen to the characters next, it is anticipation.

Increasing levels of each of the following increase the level of dramatic tension:

- **Uncertainty:** How uncertain are the characters involved about what will happen next? Put yourself in the characters shoes; judge the change in the tension based on how the characters perceive the situation.

- **Significance:** How significant are the consequences of what will happen to the central characters of the story?

An Example: Take a dramatic moment in a story such as a character that needs to walk along a dangerous cliff path. When the character first realises they will encounter danger the tension will rise, then tension will increase further. Other details such as falling rocks or slips will increase the tension further to a peak. When the cliff edge has been navigated safely the tension will drop. The pattern will be the same with a dramatic event such as a fight, argument, accident, romantic moment, where the tension will rise to a peak and then fall away as the tension is resolved.

You will be presented with one sentence at a time. Once you have read the sentence, you will press one of five keys to judge the increase or decrease in dramatic tension that this sentence caused. You will use five levels (with keyboard shortcuts in brackets):

- **Big Decrease (A):** A sudden decrease in dramatic tension of the situation. In the cliff example the person reaching the other side safely.

- **Decrease (S):** A slow decrease in the level of tension, a more gradual drop. For example the cliff walker sees an easier route out.

Annotation	Sentence
NA	Clancy Marguerian, 154, private first class of the 150 + army , sits in his foxhole.
Increase	Tired cold, wet and hungry, the only thing preventing him from laying down his rifle and walking towards the enemy lines in surrender is the knowledge that however bad he has it here, life as a 50 - 100 POW is surely much worse .
Increase	He's fighting to keep his eyes open and his rifle ready when the mortar shells start landing near him.
Same	He hunkers lower.
Increase	After a few minutes under the barrage, Marguerian hears hurried footsteps, a grunt, and a thud as a soldier leaps into the foxhole.
Same	The man's uniform is tan , he must be a 50 - 100 .
Big Increase	The two men snarl and grab at each other , grappling in the small foxhole .
Same	Abruptly, their faces come together.
Decrease	"Clancy?"
Decrease	"Rob?"
Big Decrease	Rob Hall, 97, Corporal in the 50 - 100 army grins, as the situation turns from life or death struggle, to a meeting of two college friends.
Decrease	He lets go of Marguerian's collar.
Same	" Holy shit Clancy , you're the last person I expected to see here "
Same	" Yeah " " Shit man , I didn't think I'd ever see Mr. volunteers every saturday morning at the food shelf' , not after The Reorganization at least "
Same	"Yeah Rob , it is something isn't it "
Decrease	" Man , I'm sorry, I tried to kill you there".

Table 5: One of the training annotation examples given to Mechanical Turk workers. The annotation labels are the recommended labels. This is an extract from a validation set WritingPrompts story.

- **Same (Space):** Stays at a similar level. In the cliff example an ongoing description of the event.
- **Increase (K):** A gradual increase in the tension. Loose rocks fall nearby the cliff walker.
- **Big Increase (L):** A more sudden dramatic increase such as an argument. The cliff walker suddenly slips and falls.

POST ACTUAL INSTRUCTIONS In addition to the suspense annotation. The following review questions were asked:

- Please write a summary of the story in one or two sentences.
- Do you think the story is interesting or not? And why? One or two sentences.
- How interesting is the story? 1–5

The main purpose of this was to test if the MTurk Annotators were comprehending the stories and not trying to cheat by skipping over. Some further work through can be done to tie these into the suspense measures and also the WritingPrompts prompts.

C Writing Prompts Examples

The numbers are from the full WritingPrompts test set. Since random sampling was done from these from for evaluation the numbers are not in a contiguous block. There are a couple of nonsense sentences or entirely punctuation sentences. In the model these are excluded in pre-processing but included here to match the sentence segmentation. Also there are some unusual break such as "should n't", this is because the word segmentation produced by the Spacy tokenizer.

C.1 Story 27

This is Story 27 from the test set in Figure 4, it is the same as the example in the main text:

0. As I finished up my research on Alligator breeding habits for a story I was tasked with writing , a bell began to ring loudly throughout the office .

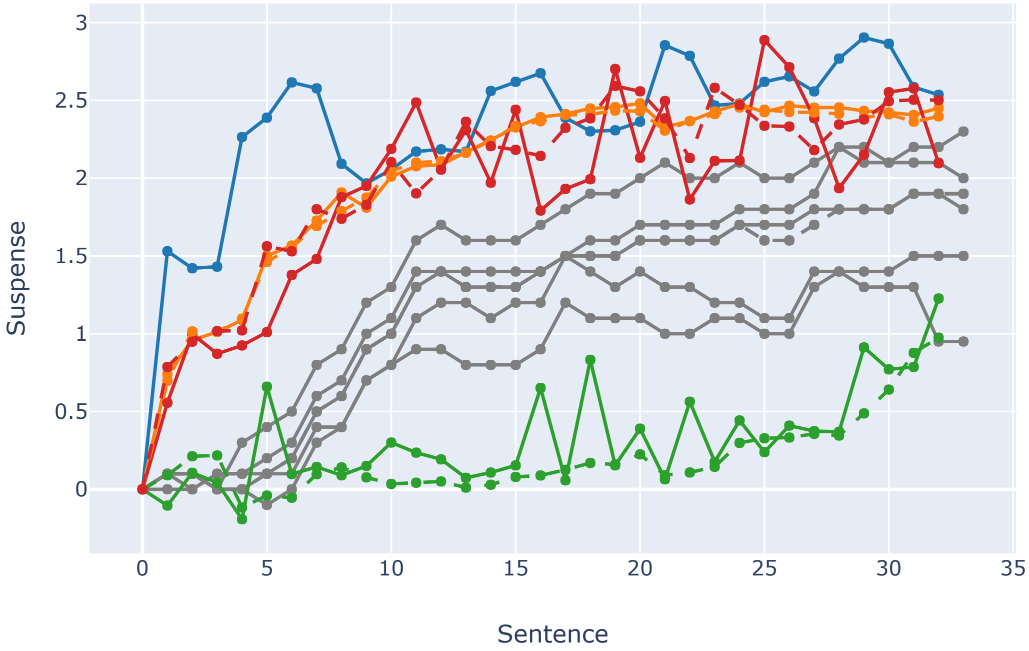

Figure 4: Story 27, **Human**, S^{Hale}, S^{Ely}, U^{Ely}, $U^{\alpha\text{Ely}}$

1. I could feel the sound vibrating off the cubicle walls .

2. I looked over my cubicle wall to ask a co - worker what the bell was for .

3. I watched as he calmly opened his desk drawer , to reveal a small armory .

4. There were multiple handguns , knives and magazines and other assorted weapons neatly stashed away .

5. “ What the hell is that for ? ”

6. I questioned loudly , and nervously .

7. The man looked me in the eyes , and pointed his handgun at my face .

8. I saw my life flash before my eyes , and could n’t understand what circumstances had arisen to put me in this position .

9. I heard the gun fire , and the sound of the shot rang through my ears .

10. I heard something hit the ground loudly behind me .

11. I turned to see the woman who had hired me yesterday , lying in a pool of blood on the floor .

12. She was holding a rifle in her arms .

13. I looked back at the man who had apparently just saved my life .

14. He seemed to be about 40 or so , well built , muscular and had a scar down the right side of his face that went from his forehead down to his beard .

15. “ She liked to go after the new hires ” he explained in a deep voice .

16. “ She hires the ones she wants to kill ”

17. I was n’t sure what to make of this , but my thoughts were cut off by the sounds of screaming throughout the building .

18. “ What ’s happening ”

19. I asked , barely able to look my savior in the eyes .

20. “ You survive today , and you ’ll receive a bonus of $5,000 and your salary will be raised 5 % ”

21. I cut the man off .

22. “ What does that ? ”

23. He continued to speak , while motioning me to stop taking .

24. “ I ’ll keep you alive , if you give me your bonus and half your raise

25. He finished .

26. I just nodded , still unable to understand the position I was in .

27. He grabbed my arm so hard I thought it would break , and pulled me over the cubicle wall , and under his desk .

28. Then , he placed a gun in my hand .

29. “ The safety is on , and it ’s fully loaded with one in the chamber ”

30. He said , pointing to the safety switch .

31. The weapon felt heavy in my hand , I flicked the safety off with my thumb and gripped the gun tightly .

32. The man looked down at his watch .

33. “ 45 minutes to go ”

C.2 Story 2066

This is Story 2066 from the test set in Figure 5:

0. The life pods are designed so we ca n’t steer .

1. Meant for being stranded in space , it broadcasts an S.O.S .

2. to the entire human empire even as it leaves the mother ship .

3. Within minutes any occupant will be gassed so they wo n’t suffer the long months , and perhaps years before a rescue .

4. As soon as your vitals show you ’re in deep sleep , it puts the entire interior into a cryogenic freeze .

5. The technology is effective , efficient and brilliant .

6. But as I ’ m being launched out of our vessel I ca n’t help but slam the hatch with my fists .

7. My ears are still ringing with the endless boom of explosions and my eyes covered in blind spots from the flashes .

8. The battle had been swift , and we humans had lost .

9. Captain ’s orders : Abandon ship .

10. Which was why I was stuck here , counting the seconds before I got put into stasis .

11. This was no Titanic .

12. There were ample pods for the entire crew , by the time the call was made only half of us had access to the escape pods , and a quarter of those were injured , a condition that no matter how advanced our technology was , made the life pod a null option .

13. No use being cryogenically frozen if you bleed out before the temperature even drops .

14. Better men and women than I were stuck alive on the ship , and I had to abandon them to whatever their fate may be .

15. I sit back and harness myself into the chair .

16. No use getting worked up over survivor ’s guilt now .

17. I ’ll do that when I thaw .

18. *

19. * *

20. The first thing I notice is the cold .

21. I ’ m too cold .

22. I shiver , my uniform plastered to me .

23. I frown at its tattered appearance .

24. What had happened ?

25. The last thing I remember is ...

26. The life pod .

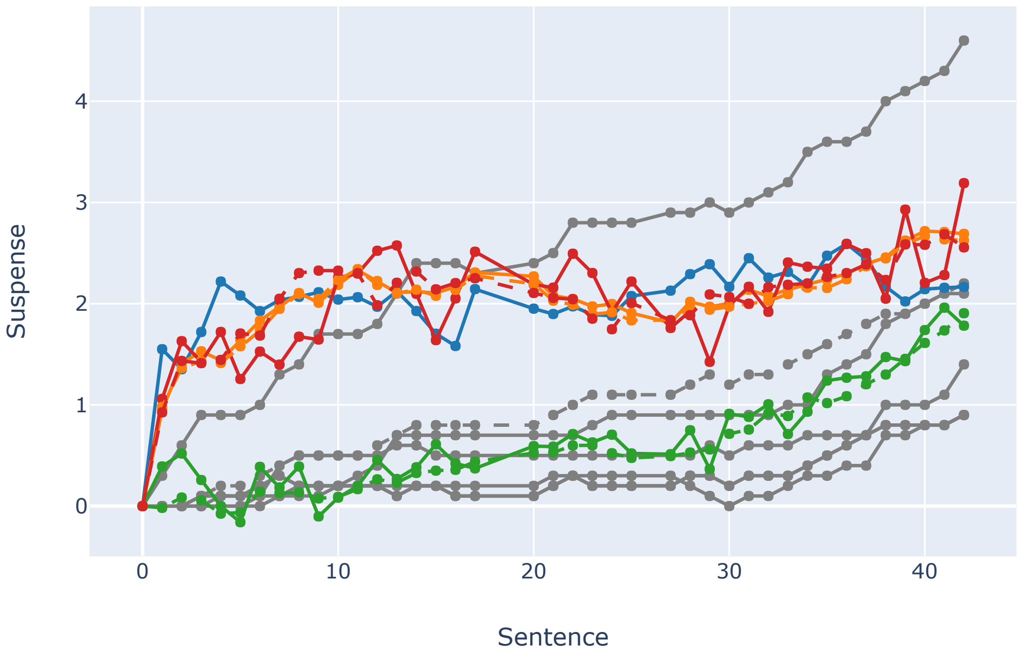

Figure 5: Story 2066, **Human**, S^{Hale}, S^{Ely}, U^{Ely}, $U^{\alpha\text{Ely}}$

27. I ’ m still in it .

28. But I ’ ve been picked up .

29. Someone on the outside has initiated the thaw cycle .

30. At once I ’ m struck by relief .

31. Then anxiety .

32. How long was I out ?

33. How many of the crew survived ?

34. Their screams are coming back to me now , and I squirm with the pain .

35. “ Please do n’t let me be the only one , ” I whisper to myself , half pleading with fate , half praying to a God .

36. The hatch swings open .

37. The lump in my throat drops to my toes with the weight of lead .

38. A gun greets me .

39. Slowly , I put my hands behind my head .

40. There ’s no mistaking the alien wielding it .

41. The brute features are familiar , too familiar .

42. I ’ ve been rescued by the wrong side .

C.3 Story 3203

This is Story 3203 from the test set in Figure 6:

0. I swore never to kill .

1. I swore that I will never stoop down to their level .

2. That we , the guardians of justice , can and will achieve our goals through the peaceful way .

3. But as I stood there , at the edge of the cliff , staring at the hideous smile that has tormented me for far too long , I could feel my vow slowly breaking before me .

4. “ So what it ’s gon na be Batsy ?

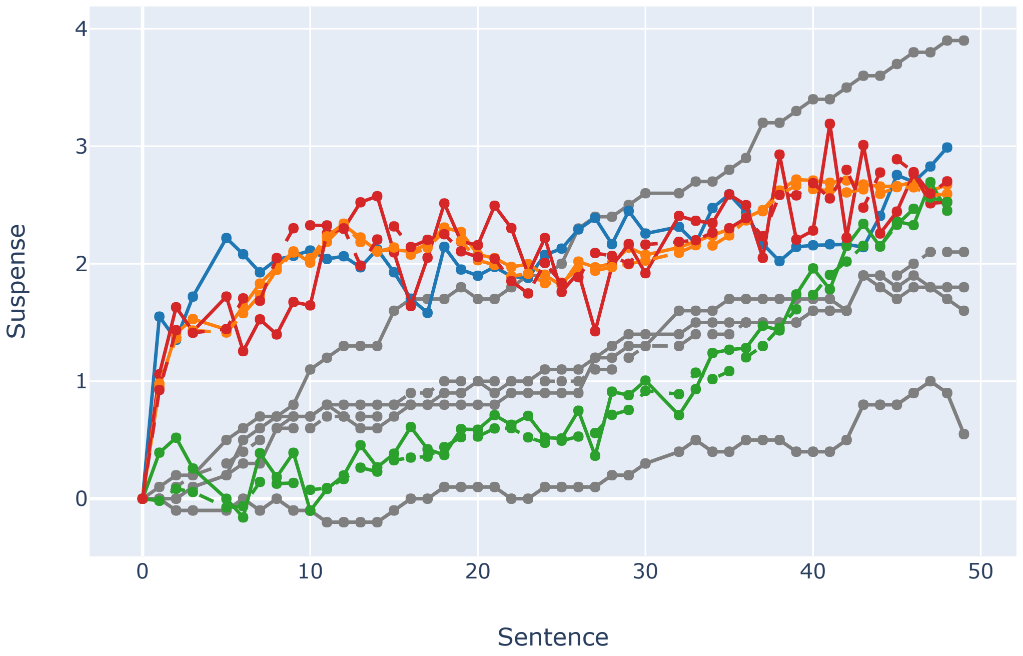

Figure 6: Story 3203, **Human**, S^{Hale}, S^{Ely}, U^{Ely}, $U^{\alpha\text{Ely}}$

5. Will you choose to kill the evil crazy clown , or are you going to let poor Miss Lane fall to her death ?

6. Tick tock tick tock , time 's ticking ! "

7. I gritted my teeth .

8. Lois was suspended in mid - air , 12 stories high , her life hanging by the mere minutes .

9. Around me , the League lay incapacitated , having fallen to Joker 's devious ambush .

10. I turned towards Clark , hoping that he would have woken up by now .

11. No luck .

12. The Kryptonite knock out gas had worked its miracle .

13. As fate would have it , only two of us are left .

14. Two bitter rivals to the very end .

15. " Let her go , Joker !

16. This fight is between you and me ! "

17. I shouted .

18. My mind raced for possible solutions .

19. A well - aimed batarang could free Lois , but I have to rappel to her in time .

20. Too risky with Joker free .

21. I could try knocking him out , but that would not leave me enough time to- " Tsk tsk tsk , my dear Bats .

22. Trying to stall for time , are n't you ?

23. How many times must I tell you that it wo n't work !

24. I know you , Bats , better than you know yourself .

25. In fact ... "

26. He took out a remote , and pushed one of the bright red buttons on it .

27. The cable jerked downwards , closer to the barrel of Joker venom .
28. “ ... for every minute you spend thinking , Miss Lane will be closer to smiley face land .
29. How about that !
30. Hahahaha ! ”
31. It was right then when I lost it .
32. I leaped from my spot , headed straight for the Prince of Clowns .
33. I thought about the last time we almost lost Lois .
34. Clark was so close to unleashing a destructive rampage across Metropolis .
35. Too close .
36. And it was on that day when every member of the League swore an oath to protect Lois no matter what it takes , no matter what the cost , even if it meant breaking our own sacred vows .
37. Superman was too great an asset to be lost .
38. Joker knew that .
39. From the very moment he saw the destruction Clark unleashed .
40. And he has been targeting Lois ever since .
41. The blade plunged through his chest and into his heart surprisingly quick .
42. I had expected the Joker to have a fail safe mechanism , but apparently he did not .
43. He wanted me to do it .
44. The blood splattered against my suit , as the sickening sound of flesh tearing apart filled my ears .
45. And as all these happened , the Joker kept laughing , his hysterical voice filling the air .
46. He laughed and laughed , until his voice gradually grew weaker , softer .
47. Before he drew his last breath , he raised his bloodied left hand and patted me on my cowl .
48. “ Hehehe ... I win , Batsy . ”

D Turning Points Examples

This section is the full text output with some example plots from Turning Points TRIPOD dataset.

D.1 15 Minutes

The full text for the synopsis of 15 Minutes in Figure 7, this is the same example as is given in the main text:

0. After getting out of prison , ex - convicts Emil Slovak (Karel Roden) and Oleg Razgul (Oleg Taktarov) travel to New York City to meet a contact in order to claim their part of a bank heist in
1. Russia (or somewhere in the Czech Republic) .
2. Within minutes of arriving , Oleg steals a video camera .
3. They go to the brownstone apartment of their old partner Milos Karlova (Vladimir Mashkov) and his wife Tamina , and demand their share .
4. When Milos admits that he spent it , an enraged Emil kills him with a kitchen knife , then breaks Tamina ’s neck as Oleg tapes it with his new camera .
5. The couple ’s neighbor , Daphne Handlova (Vera Farmiga) , witnesses everything , but she escapes before they can get to her .
6. To cover up the crime , they douse the bodies in acetone , carefully position them on the bed , and burn down the apartment , intending to pass it off as an accident .
7. Jordy Warsaw (Edward Burns) , an arson investigator , and NYPD detective Eddie Flemming (Robert De Niro) are called to the scene .
8. Flemming is a high profile detective who frequently appears on the local tabloid TV show Top Story .
9. Flemming and Warsaw decide to work the case together .
10. They eventually determine that Milos was stabbed so hard that the knife ’s tip broke off and lodged in his spine .

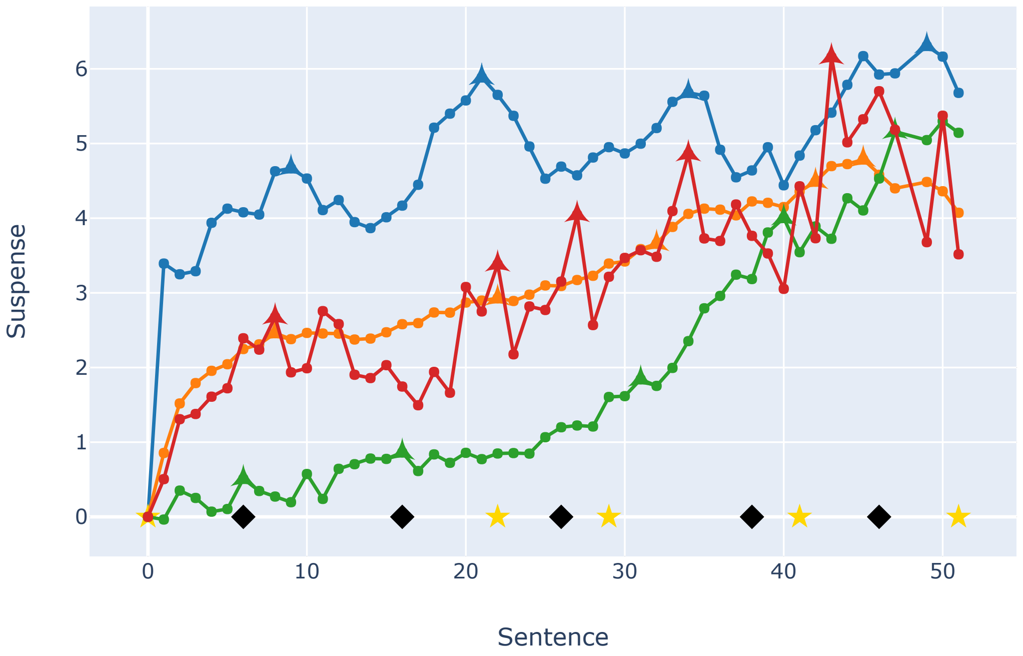

Figure 7: The film *15 Minutes*, S^{Hale}, S^{Ely}, U^{Ely}, $U^{\alpha\text{Ely}}$,◆ theory baseline, ★ TP annotations, triangles are predicted TPs.

11. While checking out the crowd outside , Warsaw spots Daphne trying to get his attention .

12. When he finally gets to where she was , she is gone , but Warsaw manages to produce a sketch of the witness .

13. Emil , who got hold of Daphne 's wallet when she fled the apartment earlier , realizes that Daphne is in the country illegally and will be deported if she calls the police .

14. He contacts an escort service from a business card he found in Daphne 's wallet .

15. He asks for a Czech girl hoping she will arrive .

16. When Honey , a regular call girl , arrives instead , he stabs and kills her , but not before getting the address of the escort service from her .

17. Oleg tapes the entire murder .

18. In fact , he tapes everything he can ; a wannabe filmmaker , he aspires to be the next Frank Capra .

19. Flemming and Warsaw investigate her murder , determine the link to the fire , and also visit the escort service .

20. Rose Heam (Charlize Theron) runs the service and tells them that the girl they are looking for (Daphne) does not work for her but rather a local hairdresser , and she just told the same thing to

21. a couple other guys that were asking the same questions .

22. Flemming and Warsaw then rush to the hairdresser but get there just after Emil and Oleg warn the girl not to say anything to anyone .

23. As Flemming puts Daphne into his squad car , he notices Oleg taping them from across the street .

24. A foot chase begins , culminating in Flemming 's partner getting shot and his wallet stolen .

25. Emil finds a card with Flemming 's name and address in it .

26. He gets very jealous of Flemming 's celebrity status and is convinced that anyone in America can do whatever they want and get away with it .

27. On the night that Flemming is to propose to his girlfriend Nicolette Karas (Melina Kanakaredes) , Oleg and Emil sneak into his house and knocks him unconscious , later taping him to a chair .

28. While Oleg is recording , Emil explains his plan - he will kill Flemming , then he will sell the tape to Top Story , and when he is arrested , he will plead insanity .

29. After being committed to an insane asylum he will declare that he is actually sane .

30. Because of double jeopardy , he will get off , collecting the royalties from his books and movies .

31. Flemming starts attacking them with his chair (while still taped to it) and almost gets them but Emil stabs him in the abdomen , and putting a pillow on Flemming , killing him .

32. The entire city is in mourning and Emil calls Robert Hawkins (Kelsey Grammer) , the host of Top Story , to tell him he has a tape of the killing and is willing to sell it .

33. Robert pays him a million dollars for the tape .

34. Warsaw and the entire police force are furious with Robert and can not believe he would air it , especially since his main reporter is Nicolette .

35. At the same time , Emil and Oleg try to kill Warsaw and Daphne by booby - trapping Daphne 's apartment .

36. The two narrowly escape the resulting fire .

37. On the night it is aired Emil and Oleg sit in a Planet Hollywood to watch it with the rest of the public .

38. As the clip progresses , the customers react with horror at the brutality of it , and a few begin to notice Emil and Oleg are right there with them , Oleg actually smiling at the results of his work , and panic takes place .

39. Emil explains his betrayal to Oleg and as he about to execute Emil with a gun , Oleg stabs him in the arm .

40. The police come in and arrest the wounded Emil , while Oleg escapes .

41. They put Emil in Warsaw 's squad car but instead of taking him to the police station , Warsaw takes him to an abandoned warehouse where he is going to kill him .

42. The police arrive just in time and take Emil away .

43. Everything goes as planned as Emil is now a celebrity and is pleading insanity .

44. His lawyer agrees to work for 30

45. Meanwhile , Oleg is jealous of the notoriety that Emil is receiving .

46. While being led away with his lawyer and all the media , Warsaw gets into an argument with the lawyer while the Top Story crew is taping the whole thing .

47. Oleg gives Hawkins the part of the tape where Emil explains his plan to Flemming , proving he was sane the whole time (Oleg presumably kept this part of the tape on hand as part of an " insurance policy "") ."

48. Hawkins shouts out to Emil and explains to him the evidence he now has .

49. Emil pushes a policeman down , takes his gun and shoots Oleg .

50. Emil grabs Flemming 's fiancĂŠe , who is covering the news story , and threatens to shoot her .

51. He is finally cornered by the police and Warsaw .

52. Against orders , Warsaw shoots Emil a dozen times in the chest in order to avenge Eddie 's death .

53. An officer shouts that Oleg is still alive , and Hawkins rushes to him to get footage just as Oleg says the final few words to his movie he is taping just before he dies (with the Statue of Liberty in the background) .

54. Shortly afterward , Hawkins approaches Warsaw and tries to cultivate the same sort of arrangement he had with Flemming , suggesting the power an arrangement would give him .

55. In response , Warsaw punches out Hawkins and leaves the scene as the police officers smile in approval .

D.2 Pretty Woman

The full text for the synopsis of the film Pretty Woman in Figure 8:

0. Edward Lewis (Gere), a successful businessman and "corporate raider", takes a detour on Hollywood Boulevard to ask for directions. Receiving little help, he encounters a prostitute named Vivian Ward (Roberts) who is willing to assist him in getting to his destination.

1. The morning after, Edward hires Vivian to stay with him for a week as an escort for social events.

2. Vivian advises him that it "will cost him," and Edward agrees to give her $3,000 and access to his credit cards.

3. Vivian then goes shopping on Rodeo Drive, only to be snubbed by saleswomen who disdain her because of her unsophisticated appearance.

4. Initially, hotel manager Barnard Thompson (Hector Elizondo) is also somewhat taken aback.

5. But he relents and decides to help her buy a dress, even coaching her on dinner etiquette.

6. Edward returns and is visibly amazed by Vivian's transformation. The business dinner does not end well, however, with Edward making clear his intention to dismantle Morse's corporation once it was bought, close down the shipyard which Morse spent 40 years building, and sell the land for real estate.

7. Morse and his grandson abandon their dinner in anger, while Edward remains preoccupied with the deal afterward.

8. Back at the hotel, Edward reveals to Vivian that he had not spoken to his recently deceased father for 14 and half years.

9. Later that night, the two make love on the grand piano in the hotel lounge.

10. The next morning, Vivian tells Edward about the snubbing that took place the day before.

11. Edward takes Vivian on a shopping spree.

12. Vivian then returns, carrying all the bags, to the shop that had snubbed her, telling the salesgirls they had made a big mistake.

13. The following day, Edward takes Vivian to a polo match where he is interested in networking for his business deal.

14. While Vivian chats with David Morse, the grandson of the man involved in Edward's latest deal, Philip Stuckey (Edward's attorney) wonders if she is a spy.

15. Edward re-assures him by telling him how they met, and Philip (Jason Alexander) then approaches Vivian and offers to hire her once she is finished with Edward, inadvertently insulting her.

16. When they return to the hotel, she is furious with Edward for telling Phillip about her.

17. She plans to leave, but he apologizes and persuades her to see out the week.

18. Edward leaves work early the next day and takes a breath-taking Vivian on a date to the opera in San Francisco in his private jet. She is clearly moved by the opera (which is La Traviata, whose plot deals with a rich man tragically falling in love with a courtesan).

19. While playing chess with Edward after returning, Vivian persuades him to take the next day off.

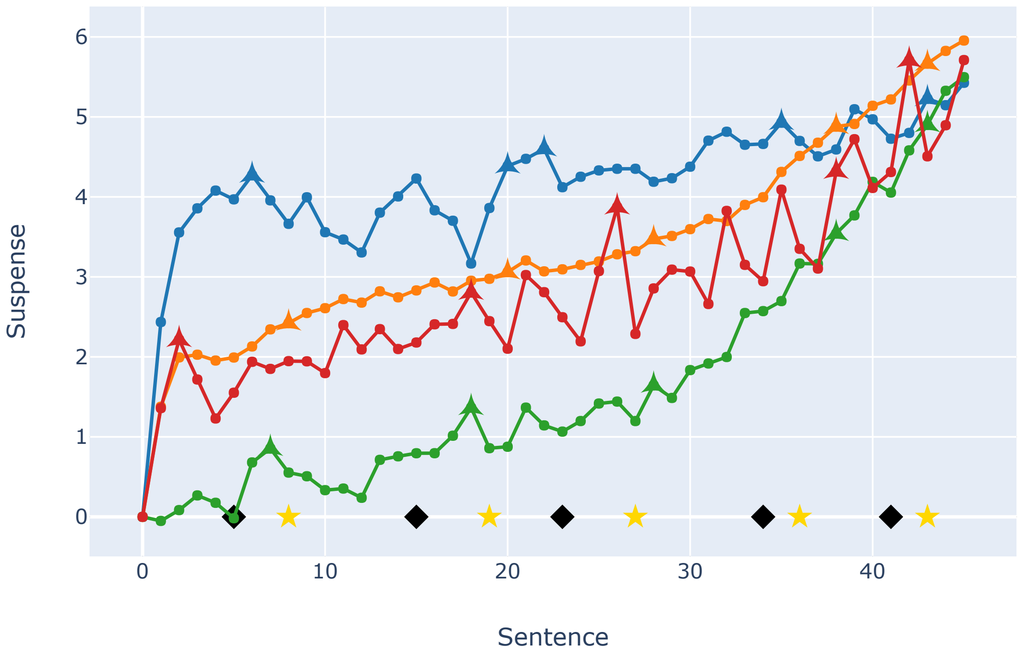

Figure 8: The film Pretty Woman, S^{Hale}, S^{Ely}, U^{Ely}, $U^{\alpha\text{Ely}}$, ◆ theory baseline, ★ TP annotations, triangles are predicted TPs.

20. They spend the entire day together, and then have sex, in a personal rather than professional way.

21. Just before she falls asleep, Vivian admits that she’s in love with Edward.

22. Over breakfast, Edward offers to put Vivian up in an apartment so he can continue seeing her.

23. She feels insulted and says this is not the ”fairy tale” she wants.

24. He then goes off to work without resolving the situation.

25. Vivian’s friend, Kit De Luca (Laura San Giacomo), comes to the hotel and realizes that Vivian is in love with Edward.

26. Edward meets with Mr. Morse, about to close the deal, and changes his mind at the last minute.

27. His time with Vivian has shown him another way of living and working, taking time off and enjoying activities for which he initially had little time.

28. As a result, his strong interest towards his business is put aside.

29. He decides that he would rather help Morse than take over his company.

30. Furious, Philip goes to the hotel to confront Edward, but only finds Vivian there.

31. He blames her for changing Edward and tries to rape her.

32. Edward arrives in time to stop Philip, chastising him for his greed and ordering him to leave the room.

33. Edward tends to Vivian and tries to persuade her to stay with him because she wants to, not because he’s paying her.

34. She refuses once again and returns to the apartment she shares with Kit, preparing to leave for San Francisco to earn a G.E.D. in the hopes of a better life.

35. Edward gets into the car with the chauffeur that took her home.

36. Instead of going to the airport, he goes to her apartment arriving accompanied by music from La Traviata.

37. He climbs up the fire escape, despite his fear of heights, with a bouquet of roses clutched between his teeth, to woo her.

38. His leaping from the white limousine, and then climbing the outside ladder and steps, is a visual urban metaphor for the knight on white horse rescuing the ”princess” from the tower, a childhood fantasy Vivian told him about.

39. The film ends as the two of them kiss on the fire escape.

D.3 Slumdog Millionaire

The full text for the synopsis of the film Slumdog Millionaire, in Figure 9:

0. In Mumbai in 2006, eighteen-year-old Jamal Malik (Dev Patel), a former street child (child Ayush Mahesh Khedekar, adolescent Tanay Chheda) from the Juhu slum, is a contestant on the Indian version of Who Wants to Be a Millionaire?, and is one question away from the grand prize.

1. However, before the Rs.

2. 20 million question, he is detained and interrogated by the police, who suspect him of cheating because of the impossibility of a simple ”slumdog” with very little education knowing all the answers.

3. Jamal recounts, through flashbacks, the incidents in his life which provided him with each answer.

4. These flashbacks tell the story of Jamal, his brother Salim (adult Madhur Mittal, adolescent Ashutosh Lobo Gajiwala, child Azharuddin Mohammed Ismail), and Latika (adult Freida Pinto, adolescent Tanvi Ganesh Lonkar, child Rubina Ali).

5. In each flashback Jamal has a point to remember one person, or song, or different things that lead to the right answer of one of the questions.

6. The row of questions does not correspond chronologically to Jamal’s life, so the story switches between different periods (childhood, adolescence) of Jamal.

7. Some questions do not refer to points of his life (cricket champion), but by witness he comes to the right answer.

8. Jamal’s flashbacks begin with his managing, at age five, to obtain the autograph of Bollywood star Amitabh Bachchan, which his brother then sells, followed immediately by the death of his mother during the Bombay Riots.

9. As they flee the riot, they run into a child version of the God Rama, Salim and Jamal then meet Latika, another child from their slum.

10. Salim is reluctant to take her in, but Jamal suggests that she could be the third musketeer, a character from the Alexandre Dumas novel (which they had been studying — albeit not very diligently — in school), whose name they do not know.

11. The three are found by Maman (Ankur Vikal), a gangster who tricks and then trains street children into becoming beggars.

12. When Jamal, Salim, and Latika learn Maman is blinding children in order to make them more effective as singing beggars, they flee by jumping onto a departing train.

13. Latika catches up and takes Salim’s hand, but Salim purposely lets go, and she is recaptured by the gangsters.

14. Over the next few years, Salim and Jamal make a living travelling on top of trains, selling goods, picking pockets, working as dish washers, and pretending to be tour guides at the Taj Mahal, where they steal people’s shoes.

15. At Jamal’s insistence, they return to Mumbai to find Latika, discovering from Arvind, one of the singing beggars, that she has been raised by Maman to become a prostitute and that her virginity is expected to fetch a high price.

16. The brothers rescue her, and Salim draws a gun and kills Maman.

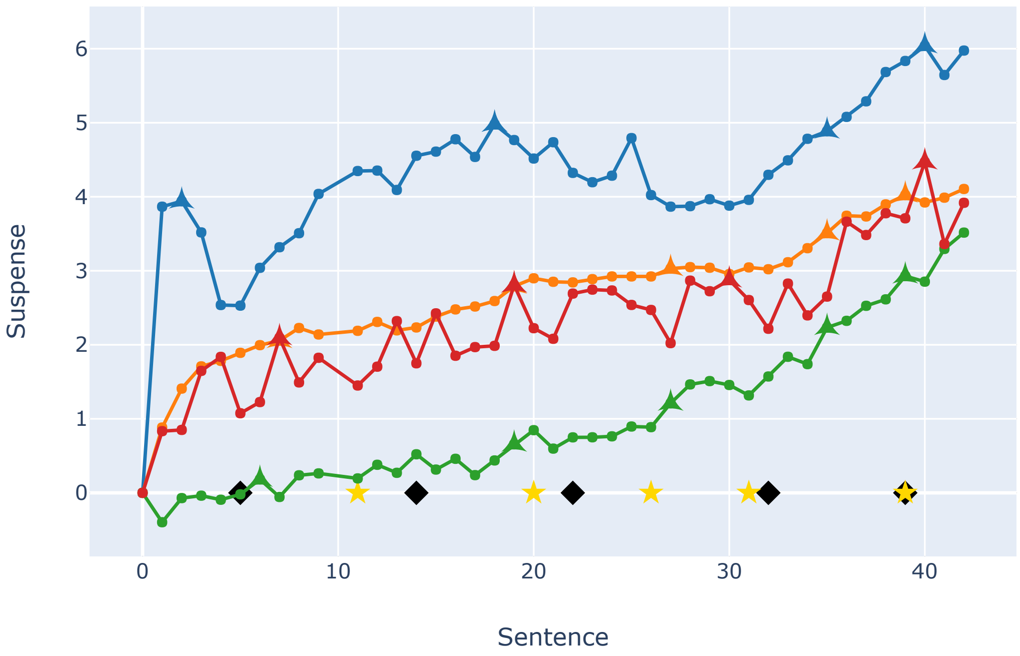

Figure 9: Slumdog Millionare, $S^{\textbf{Hale}}$, $S^{\textbf{Ely}}$, $U^{\textbf{Ely}}$, $U^{\alpha\textbf{Ely}}$, ◆ theory baseline, ★ TP annotations, triangles are predicted TPs.

17. Salim then manages to get a job with Javed (Mahesh Manjrekar), Maman's rival crime lord.

18. Arriving at their hotel room, Salim orders Jamal to leave him and Latika alone.

19. When Jamal refuses, Salim draws a gun on him, and Jamal leaves after Latika persuades him to go away (presumably so he wouldn't get hurt by Salim).

20. Years later, while working as a tea server at an Indian call centre, Jamal searches the centre's database for Salim and Latika.

21. He fails in finding Latika but succeeds in finding Salim, who is now a high-ranking lieutenant in Javed's organization, and they reunite.

22. Salim is regretful for his past actions and only pleads for forgiveness when Jamal physically attacks him.

23. Jamal then bluffs his way into Javed's residence and reunites with Latika.

24. While Jamal professes his love for her, Latika asks him to forget about her.

25. Jamal promises to wait for her every day at 5 o'clock at the VT station.

26. Latika attempts to rendezvous with him, but she is recaptured by Javed's men, led by Salim.

27. Jamal loses contact with Latika when Javed moves to another house, outside of Mumbai.

28. Knowing that Latika watches it regularly, Jamal attempts to make contact with her again by becoming a contestant on the show Who Wants to Be a Millionaire?

29. He makes it to the final question, despite the hostile attitude of the show's host, Prem Kumar (Anil Kapoor), and becomes a wonder across India.

30. Kumar feeds Jamal the incorrect response to the penultimate question and, when Jamal still gets it right, turns him into the police on suspicion of cheating.

31. Back in the interrogation room, the police inspector (Irrfan Khan) calls Jamal's explanation ”bizarrely plausible”, but thinks he is not a liar and, ripping up the arrest warrant, allows him to return to the show.

32. At Javed's safehouse, Latika watches the news coverage of Jamal's miraculous run on the show.

33. Salim, in an effort to make amends for his past behaviour, quietly gives Latika his mobile phone and car keys, and asks her to forgive him and to go to Jamal.

34. Latika, though initially reluctant out of fear of Javed, agrees and escapes.

35. Salim fills a bathtub with cash and sits in it, waiting for the death he knows will come when Javed discovers what he has done.

36. Jamal's final question is, by coincidence, the name of the third musketeer in The Three Musketeers, a fact he never learned.

37. Jamal uses his Phone-A-Friend lifeline to call Salim's cell, as it is the only phone number he knows.

38. Latika succeeds in answering the phone just in the nick of time, and, while she does not know the answer, tells Jamal that she is safe.

39. Relieved, Jamal randomly picks Aramis, the right answer, and wins the grand prize.

40. Simultaneously, Javed discovers that Salim has helped Latika escape after he hears Latika on the show.

41. He and his men break down the bathroom door, and Salim kills Javed, before being gunned down himself at the hands of Javed's men.

42. With his dying breath, Salim gasps, ”God is great.”

43. Later that night, Jamal and Latika meet at the railway station and kiss.

44. The movie ends with a dance scene on the platform to ”Jai Ho”.