

Story Comprehension for Predicting What Happens Next

Snigdha Chaturvedi Haoruo Peng Dan Roth

University of Illinois, Urbana-Champaign

{snigdha,hpeng7,danr}@illinois.edu

Abstract

Automatic story comprehension is a fundamental challenge in Natural Language Understanding, and can enable computers to learn about social norms, human behavior and commonsense. In this paper, we present a story comprehension model that explores three distinct semantic aspects: (i) the sequence of events described in the story, (ii) its emotional trajectory, and (iii) its plot consistency. We judge the model’s understanding of real-world stories by inquiring if, like humans, it can develop an expectation of what will happen next in a given story. Specifically, we use it to predict the correct ending of a given short story from possible alternatives. The model uses a hidden variable to weigh the semantic aspects in the context of the story. Our experiments demonstrate the potential of our approach to characterize these semantic aspects, and the strength of the hidden variable based approach. The model outperforms the state-of-the-art approaches and achieves best results on a publicly available dataset.

1 Introduction

Narratives are a fundamental part of human language and culture. They serve as vehicles to share experiences, information and goals. For these reasons, automatically understanding stories is an interesting but challenging task for Computational Linguists (Mani, 2012). Story comprehension involves not only an array of NLP capabilities, but also some common sense knowledge and an understanding of normative social behavior (Charniak, 1972). Past research has focused on various aspects of story understand-

Context: One day Wesley’s auntie came over to visit. He was happy to see her, because he liked to play with her. When she started to give his little sister attention, he got jealous. He got angry at his auntie and bit her hand when she wasn’t looking.

Incorrect Ending: She gave him a cookie for being so nice.

Correct Ending: He was scolded.

Figure 1: Example from the story-cloze task: predict the correct ending to a given short story out of provided options.

ing such as identifying character personas (Bamman et al., 2014; Valls-Vargas et al., 2015), interpersonal relationships (Chaturvedi, 2016), plot-patterns (Jockers, 2013), narrative structures (Finlayson, 2012). There has also been an interest in predicting what is expected to happen next in a piece of text (Chambers and Jurafsky, 2008). Human readers are good at *filling-in-the-gaps* or inferring information that is not explicitly stated in the text. However, computers are not yet able to match their performance on predicting what could be the likely next step in a given sequence of events described in a story.

Recently, Mostafazadeh et al. (2016) introduced the story-cloze task for testing this ability, albeit without the aspect of language generation. This task requires choosing the correct ending to a given four sentences long story (also referred to as *context*) from two provided alternatives. Fig. 1 shows an example story consisting of a short context, and two ending options.

In this work we address this story-cloze task. While the short nature and third person narrative style of these stories help us circumvent the problem of speaker identification and processing long

dialogues, the crowdsourced dataset ensures that they reflect real-world and commonsense stories. Our approach emphasizes the joint contribution of multiple aspects to story understanding, which future research can build upon.

In this paper we explore three semantic aspects of story understanding: (i) the sequence of events described in the story, (ii) the evolution of sentiment and emotional trajectories, and (iii) topical consistency. The first aspect is motivated from approaches in semantic script induction, and evaluates if events described in an ending-alternative *are likely* to occur within the sequence of events described in the preceding context. For example, in the story in Fig. 1, Wesley gets angry and bites his sister’s hand. So, a next likely step might suggest that he would be scolded. However, there are multiple semantic aspects to story understanding beyond analyzing events and scripts. Stories often describe characters (e.g. Wesley) who need to be viewed as social and emotional agents. They not only describe events involving these characters, but also reflect their social lives and emotional states. Our model captures this by evaluating if the sentiment described in an ending option *makes sense* considering the context of the story. For example, in the story in Fig. 1, the general sentiment of being *scolded* is better aligned with the sentiment of Wesley being *angry* and *jealous*, compared to that of *being nice*. Also, stories generally revolve around coherent themes and topics. Our model accounts for that by analyzing if the topic of an ending option is consistent with the preceding context. We present a log-linear model that is used to weigh the various aspects of the story using a hidden variable. It then uses this hidden variable to predict the correct ending for the given story.

We demonstrate the strength of our approach by comparing it with the existing state-of-the-art methods for this task. We first validate the predictive potential of the features that correspond to the three semantic aspects through a simple classifier trained using these features. We then demonstrate the benefit of using our hidden variable approach by showing that it significantly outperforms the above mentioned classifier and other baselines, and achieves an accuracy of 77.60% on the task. Our key contributions are:

- We model story understanding as a joint model over multiple semantic aspects, and

utilize the idea for predicting a story’s end.

- We design linguistic features that incorporate world knowledge and narrative awareness.
- We present a hidden variable approach to weigh these aspects in a story’s context.
- We empirically demonstrate that our approach significantly outperforms state-of-the-art methods.

2 Predicting Story Ending

Given an L sentences long context, $\mathbf{c} = \langle c_1, c_2, c_3 \dots c_L \rangle$, and two ending-options, o_1 and o_2 , we aim to predict which ending option forms an inconsistent story. This is a binary classification task. We assume that the inconsistency can arise from one (or more) of certain semantic aspects. In this section, we first describe the intuition behind using these aspects and the features that we designed to capture them (Sec. 2.1). We then describe our model which uses a latent variable to weigh these aspects in light of the story, and then predicts its ending (Sec. 2.2).

2.1 Measuring Consistency

Our approach analyzes the following aspects of story understanding: Event-sequence, Sentiment-trajectory, and Topical Consistency.

Event-sequence: For a story, or any piece of text, to be coherent, it needs to describe a meaningful or ‘mutually entailing’ sequence of events (Chatman, 1980). For instance, in Figure 1 *Wesley got angry* \rightarrow *Wesley bit her hand* \rightarrow *Wesley was scolded* describes a more coherent sequence of events, as compared to *Wesley got angry* \rightarrow *Wesley bit her hand* \rightarrow *Wesley got a cookie*

Prior work in script-learning attempts to model such *prototypical* sequence of events (usually captured through verbs). For this task, we wanted to model events at an abstraction level that would be generalizable and yet semantically meaningful. Peng and Roth (2016) recently proposed a neural SemLM approach, to model such sequence of events using a language model of FrameNet (Baker et al., 1998) frames that are evoked in the given text. It represents an event using the corresponding predicate frame and its sense, obtained using a Semantic Role Labeler (Punyakanok et al., 2004). It also extends the frame definition to include explicit discourse markers (such as *but*, *and*) since they model relationships between frames. For example, in Fig-

ure 1, the SemLM representation for the last sentence of the context is ‘Get.01-and-bit.01’. Here, ‘01’ indicates specific predicate senses for verbs ‘get’ and ‘bit’ with ‘and’ being a discourse marker. Also, it produces ‘scold.01’ and ‘give.01’ for the correct and incorrect endings respectively. We train this language model using a log bilinear language model (Mnih and Hinton, 2007) on a collection of unannotated short stories (see Sec. 3.1) and also 20 years of New York Times data¹.

Given a sequence of frames evoked in the context, such a trained language model can then be used to get the conditional probabilities of the frame(s) evoked in each of the two ending-options. The option with more probable frame(s) is likely to be the appropriate ending. With this intuition in mind, for each of the two ending-option, o_i , we design features whose values are probabilities of frames evoked in that option (f_{o_i}), given the sequence of frames, $\langle f_1, f_2, \dots, f_D \rangle$, evoked in the context, $\mathbf{c} = \langle c_1, c_2, c_3 \dots c_L \rangle$. We consider increasingly longer frame-contexts for conditional probability computation, i.e. for each option, o_i , we extract the following features: $P(f_{o_i}|f_D)$, $P(f_{o_i}|f_D f_{D-1})$, \dots $P(f_{o_i}|f_D f_{D-1} \dots f_1)$. For each of these features, we additionally also include a comparative binary feature whose value is 1 if the conditional probability of one of the options (o_2) is greater than the corresponding conditional probability of the other option (o_1) (E.g. $P(o_2|f_D) > P(o_1|f_D)$), and -1 otherwise. Our preliminary experiments indicated that these features were helpful for supervised classification.

Sentiment-trajectory: As mentioned before, stories are different from objective texts such as news articles, as they additionally describe sentiments or emotions. Some stories can be categorized as happy stories while others as sad. However, most stories depict evolving sentiments in their plots as they progress (Vonnegut, 1981).

With the goal of modeling such sentiment trajectories, we assumed that a story can be divided into the following narrative-segments: a *beginning*, a *body*, a *climax*, and an *ending*. While this narrative-segmentation process warrants deeper research, in this paper we adopt a simple methodology. We treat the first sentence of the L sentences long context as the

beginning, the next $L - 2$ sentences are treated as the *body*, the last sentence of the context forms the *climax*, and the two options form the (possible) *ending*². We then assigned a positive, negative, or neutral sentiment to each segment, represented as $\mathbb{S}(\text{segment}) = \text{sign}(\text{number of positive words} - \text{number of negative words})$ in the segment. The sentiment polarity of a word was determined by a look-up from pre-trained sentiment lexica (Liu et al., 2005; Wilson et al., 2005)³. Thus, the L length context can now be viewed as a sequence of its segment’s sentiments. Lastly, we learn sentiment trajectories in form of N-gram language models from an unannotated corpus of short stories (Sec. 3.1) that learn: (i) $P(\mathbb{S}(\text{ending})|\mathbb{S}(\text{climax}), \mathbb{S}(\text{body}), \mathbb{S}(\text{beginning}))$; (ii) $P(\mathbb{S}(\text{ending})|\mathbb{S}(\text{climax}), \mathbb{S}(\text{body}))$; and (iii) $P(\mathbb{S}(\text{ending})|\mathbb{S}(\text{climax}))$.

The process described above learns typical sentiment trajectories over narrative-segments. However, it does not model a story’s *overall* sentiment (i.e. whether it is a happy or a sad story, *in general*). To capture this notion, we train another language model to learn $P(\mathbb{S}(\text{ending})|\mathbb{S}(\text{context}))$, where $\mathbb{S}(\text{context})$ is the sentiment of the full context (without segmentation).

Finally, for each ending option, we extract features whose values are the four conditional probabilities described above. As before we also consider four comparative binary features.

Topical Consistency: This aspect is motivated by the idea that stories are topically cohesive (Bamberg, 2012), and in a typical story, new topics (concepts, entities or ideas) are not introduced towards the end because it does not allow the story-writer enough narrative space and time to develop and describe them (Jovchelovitch and Bauer, 2000). We capture the notion of topic of a sentence using *topic-words* (the nouns and verbs appearing in it (Lapata and Barzilay, 2005)). For each option, we first *align* each of its topic-words with the most similar topic-word in one of the context-sentences, while defining the alignment score as this similarity value. We measure similarity between two words using the cosine similarity of their vector space representations (using

¹Owing to the large size of the training data and the fact that we abstract to the frame-semantic (and not verb) level, we cover most instances (76%) in our dataset.

²The reported segmentation process made sense from qualitative analysis on a random sample, and also led to superior performance compared to alternate strategies.

³Polarities of ‘negated’ word were reversed (determined from *neg* dependency relation in the corresponding sentence).

pretrained GloVe (Pennington et al., 2014) vectors). We then quantify the *topical-closeness* of an ending option with the context using averaged alignment score of its topic-words⁴. For each ending option, we extract one feature whose value is this *topical-closeness* with the context. As before, we also include a binary comparative feature.

2.2 Hidden Coherence Model

Sec. 2.1 described the three semantic consistency aspects and the corresponding features. We now describe our model which uses these features (represented as \vec{f}_{co} in the rest of this paper) to identify the (in)coherent ending-option. The model is also dependent on another feature set, $\vec{\phi}_{co}$, which will be discussed later in this section.

Formally, our model addresses the following binary classification problem: given the multi-sentence context, \mathbf{c} , and two ending-options, o_1 and o_2 , predict the answer, $a \in \{0, 1\}$. The correct ending for the story is o_1 when $a = 0$ and o_2 otherwise. Our training data consists of instances (context and ending options) labeled with corresponding answers a . It does not contain any other annotation (like semantic consistency aspects).

The model proceeds by assuming that there are K different semantic consistency aspects and that an ending-option can lead to an incoherent story by violating any of these aspects (our implementation uses $K = 3$ corresponding to the three aspects described in Sec. 2.1). The model achieves this by assuming that each instance belongs to a latent category, $z \in \{1, 2, 3 \dots K\}$, which advises the model on the importance of these aspects for the given instance. Using these definitions and assumptions, the probability of an answer given the context and the ending-options can be modeled as:

$$P(a|\mathbf{c}, o_1, o_2) = \sum_z^K P(z|\mathbf{c}, o_1, o_2)P(a|z, \mathbf{c}, o_1, o_2)$$

We parameterize $P(z|\mathbf{c}, o_1, o_2)$ as:

$$P(z|\mathbf{c}, o_1, o_2) = \frac{e^{-\vec{\lambda}_z \vec{\phi}_{co}}}{\sum_k e^{-\vec{\lambda}_k \vec{\phi}_{co}}}$$

⁴An alternative would be to compute similarity between averaged vector representations of the topic-words of the context and the ending-option(s). However, that assumes that a story is strictly about a single topic. Instead they reflect interplay of multiple related and ‘narrow topics’. E.g. a story describing a teacher walking in rain is about topics like ‘teacher’, ‘walk’, ‘rain’, etc. The correct ending option describes a passer-by helping the teacher. ‘passer-by’ was far from an average of all topics but close to the ‘walk’ topic.

where, $\vec{\phi}_{co}$ is the feature vector used for assigning a value to the hidden variable for an instance, and $\vec{\lambda}_z$ is the weight vector of the log-linear model for the z^{th} aspect. There are K weight vectors, one corresponding to each of the K aspects.

For predicting the answer, a , we assume that each aspect has a separate logistic-regression based prediction model parameterized as:

$$P(a|z, \mathbf{c}, o_1, o_2) = \frac{(e^{-\vec{w}_z \vec{f}_{co}^z})^{1-a}}{1 + e^{-\vec{w}_z \vec{f}_{co}^z}}$$

where \vec{f}_{co}^z is the feature vector constructed from the context and ending-options for the z^{th} aspect, and \vec{w}_z are the corresponding weights.

Training: The model parameters, \vec{w}_z and $\vec{\lambda}_z$, are learned during the training process by maximizing the log-likelihood of the data. We use Expectation-Maximization (Dempster et al., 1977) for training. During the E-step we compute the expectations for latent variable assignments using parameter values from the previous iteration as:

$$\langle z_n^k \rangle \propto \frac{e^{-\vec{\lambda}_k \vec{\phi}_{co}}}{\sum_{k'}^K e^{-\vec{\lambda}_{k'} \vec{\phi}_{co}}} P(a_n|z_n^k, \mathbf{c}_n, o_{1n}, o_{2n})$$

where, a subscript of n represents the n^{th} training instance out of a total of N instances. z_n^k represents n^{th} instance getting assigned to the k^{th} aspect, and $\langle \rangle$ denotes expected values.

In the M-step, given the expected assignments, we maximize the following expected log complete likelihood with respect to the model parameters using gradient ascent:

$$\begin{aligned} \langle L \rangle &= \sum_n^N \sum_k^K \langle z_n^k \rangle \left(\log \frac{e^{-\vec{\lambda}_k \vec{\phi}_{co}}}{\sum_{k'}^K e^{-\vec{\lambda}_{k'} \vec{\phi}_{co}}} \right. \\ &\quad \left. + \log \frac{(e^{-\vec{w}_k \vec{f}_{co}^k})^{1-a_n}}{1 + e^{-\vec{w}_k \vec{f}_{co}^k}} \right) \end{aligned}$$

Features: Our model uses two types of features: (i) for aspect-specific prediction model, \vec{f}_{co}^k , and (ii) for hidden aspect assignment, $\vec{\phi}_{co}$. The features extracted for each of the $K = 3$ aspects, \vec{f}_{co}^k , were described in Sec. 2.1. For the hidden aspect assignment, we needed features that could analyze the two options in light of the given context, and characterize the importance of various aspects for the given instance. One way to measure an aspect’s importance is by quantifying how different the two options are with respect to that aspect. The underlying assumption is that the option that

leads to an inconsistent story, by compromising on one of the aspects, would differ significantly from the other option in that aspect. We quantify an aspect’s importance using the normalized $L1$ distance between the corresponding features, \vec{f}_{co}^k , extracted for the two options in Sec. 2.1 (ignoring the comparative binary features, and normalizing by the number of features). Specifically, for an aspect k , let \vec{f}_1^k and \vec{f}_2^k (each of length n) then the corresponding ‘importance feature’ for this aspect = $|\vec{f}_1^k - \vec{f}_2^k|/n$. For example, for topical-consistency, for each option, we extracted 1 feature measuring its *topical-closeness* to the context. For $\vec{\phi}_{co}$ computation we consider the absolute difference between this value for the two options. To summarize, for each instance, we define $\vec{\phi}_{co}$ as a set of $K + 1$ features: K of these measure the importance of each of the aspects, while the last one is an additional always-one feature which captures the context-insensitive bias in the data.

3 Empirical Evaluation

In this section we describe our experiments.

3.1 Dataset

For our experiments, we have used a publicly available collection of commonsense short stories released by Mostafazadeh et al. (2016). It consists of about 100K unannotated five-sentences long stories. For collecting these stories, Amazon Mechanical Turk workers were asked to compose novel five-sentence long stories on everyday topics. They were prompted to write coherent stories with a specific beginning and ending, with *something* happening in between. This resulted in a wide variety in topics with causal and temporal links between the events described in the story. Also, the workers were asked to limit the length of individual sentences to 70 characters which yielded short and succinct sentences, and to not use informal language or quotations.

The dataset also contains an additional set of 3,742 four-sentences long stories (context) with two ending options, only one of which is correct. Each instance is annotated with this correctness information. This set was collected by asking Amazon Mechanical Turk workers to write a coherent and an incoherent ending to a given short story. The workers were asked to ensure that both the options shared at least one character from the story,

and that the options, in isolation, made sense. This resulted in non-trivial alternative endings, and was also validated by other human subjects for high quality. This set was divided by Mostafazadeh et al. (2016) into validation and test sets of 1871 instances each for the Story-Cloze Task, and were used for training and evaluating our model.

3.2 Baselines

We use the following baselines in our experiments: **DSSM:** (Mostafazadeh et al., 2016) It trains two deep neural networks (Huang et al., 2013) to project the context and the ending-options into the same vector space. Based on these vector representations, it predicts the ending-option with the largest cosine similarity with the context.

Msap: The task addressed in this paper was also a shared task for an EACL’17 workshop and this baseline (Schwartz et al., 2017) represents the best performance reported on its leaderboard (Mostafazadeh et al., 2017). It trains a logistic regression based on stylistic and language-model based features.

LR: Our next baseline is a simple logistic regression model which is agnostic to the fact that there are multiple types of aspects. Given a context and ending-options, it predicts the answer using the same features (Sec. 2.1) as the Hidden Coherence model but clubs them all into one feature-vector.

Majority Vote: This ensemble method uses the features extracted for each of the $K = 3$ aspects, to train K separate logistic regression models. It then makes a prediction by taking a majority vote of these K classifiers.

Soft Voting: This baseline also learns K different aspect-specific classifiers. However, instead of taking a majority vote, it computes a score for each option, o_i , as $\prod_k P_k(\text{ending} = o_i | \mathbf{c}, o_1, o_2)$. Here P_k represents the probability obtained from the k^{th} logistic regression. The final prediction corresponds to the option with greater score.

Aspect-aware Ensemble: Like the voting methods, this baseline also trains K different aspect-specific classifiers. However, it makes the final prediction by training another logistic regression over their predictions.

3.3 Quantitative Results

Table 1 shows accuracies of various models on the held-out test set. An always-one classifier would get 51.3% accuracy on the task and human performance is reported to be 100% (Mostafazadeh

Model	Accuracy
DSSM (Mostafazadeh et al., 2016)	58.5%
Msap (Schwartz et al., 2017)	75.2%
Majority Voting	69.5% *
Aspect aware ensemble	71.5% *
LR	74.4% *
Soft Voting	75.1%
Hidden Coherence Model	77.6% *

Table 1: Test-set accuracies of various models. Our Hidden Coherence Model outperforms competitive baselines and state-of-the-art system.

et al., 2016). A * indicates that the model’s accuracy was significantly better than the previous best model in the table (using McNemar’s test with $\alpha = 0.1$). We can see that the logistic regression, LR, outperforms the DSSM model indicating the strength of the features extracted for the various story-consistency aspects. Also, the Soft Voting approach gives us slight benefit over the LR model, possibly because of increased expressivity which includes better *organization* of features into groups or aspects. Majority Vote, in spite of sharing a similar classifier structure, does not perform as well. This might happen because it takes a *hard* vote of individual classifiers, which might be detrimental to model performance if one of the classifiers is weak. Our analysis in Sec. 3.4 shows that the topical-consistency features indeed result in a relatively weak classifier. The Aspect aware Ensemble performs better possibly because of its ability to weight the aspects (though not in context of the story).

Lastly, we can see that the proposed Hidden Coherence model, with an accuracy of 77.60%, outperforms all other models. The superior performance of our model indicates the benefit of the context-sensitive weighing of individual consistency aspects.

3.4 Ablation Study

We now investigate the predictive value of the various aspect-specific features. Table 2 shows the performance of a logistic regression model trained using all the features (All) and then using individual feature-groups. We can see that the features extracted from the aspect analyzing the event-sequence have the strongest predictive power, followed by those characterizing Sentiment-trajectory. The features measuring top-

Features	Accuracy
All	74.4%
Event-sequence	71.6%
Sentiment	64.5%
Topic	55.2%

Table 2: Performance comparison of various aspect features. Our event-sequence based features are most helpful followed by Sentiment-trajectory and then Topical Consistency based features.

ical consistency result in lowest accuracy but they still perform better than random on the task.

3.5 Qualitative Results

Table 3 shows example stories, and weights given to the three aspects. An aspect’s weight is its contribution towards the predicted output, and is shown as a bar of vertically stacked blocks in the last column. A block’s height is proportional to its aspect’s weight. Light grey block represents Event-sequence, and dark grey and black blocks represent Sentiment-trajectory and Topical consistency respectively.

The first row describes the story of a man hurting himself. A human reader can guess from commonsense knowledge that people usually recover (correct ending) after being hurt and do not repeat their mistake (incorrect ending). Accordingly, our model also primarily used the aspect analyzing events in this story, which is indicated by the long light grey block in its weight bar. Also, we can see that the topic of both the options is consistent with the story, and the model gave a very small weight to the Topical Consistency aspect indicated by the almost indiscernible black block in its weight bar. Similarly, the second row describes the story of Pam being proud of her yard work. There is a striking sentimental contrast between the two options (*upset* versus *satisfied*), and the model relies primarily on sentiments (dark grey). The last row, describes the story of Maria making candy apples. The incorrect ending introduces a new entity/idea, *apple pie*, resulting in topical incoherence of this option with the rest of the story. The model relies primarily on topic (black) and events (light grey). Reliance on events makes sense because it is likely for a person to *enjoy* what they fondly *cook*. The model gave a weight of 40% to the topical aspect, which is high as compared to its average weight across the dataset.




Context	Incorrect Ending	Correct Ending	Weights
He didn't know how the television worked. He tried to fix it, anyway. He climbed up on the roof and fiddled with the antenna. His foot slipped on the wet shingles and he went tumbling down.	He decided that was fun and to try tumbling again.	Thankfully, he recovered .	
Pam thought her front yard looked boring. So she decided to buy several plants. And she placed them in her front yard. She was proud of her work.	Pam was upset at herself.	Pam was satisfied .	
Maria smelled the fresh Autumn air and decided to celebrate. She wanted to make candy apples. She picked up the ingredients at a local market and headed home. She cooked the candy and prepared the apples.	Maria's apple pie was delicious.	She enjoyed the candy apples.	

Table 3: Examples of stories, ending-options, and aspect weights learned by our model. Aspect weights are shown as bars of stacked blocks in the last column (light grey, dark grey and black represent Event-sequence, Sentiment-trajectory and Topical Consistency respectively). A block's height is proportional to its component's weight. Black blocks are sometimes not visible because there were too small.

3.6 Discussion

Error Analysis: Table 4 shows examples of stories for which our model could not predict the correct ending. We believe that many of these stories require a deeper understanding of language and commonsense. For example, in the story described in the first row, the protagonist accepted an invitation from his friends to go to a club but danced terribly, and so he was asked to stay home the next time. To make the correct prediction in this story, the model not only needs to understand that if one does not dance well at a club they are likely to be not invited in the future, but also that staying home is the same as not getting invited. Similarly, the second row shows a story in which Johnny asks Anita out, but she makes an excuse. He later sees her with another guy and decides not to ask her out again. This example requires identifying that Anita's excuse was a lie indicating her disinterest in Johnny, which makes it unlikely for Johnny to invite her again. It also needs an understanding of inter-personal relationships, i.e. seeing a potential lover with another person leads to estrangement.

Social Analysis: To further explore the significance of social relations in stories, we consider the special case of romantic stories. We use a deterministic heuristic to identify romantic stories using lexical matches with a handcrafted list containing words like *marry*, *proposal*, *girlfriend*, *ask out*, etc. We then applied the following two rules: (i) if a story contains two characters, then out-

put the option whose sentiment matches that of the context, (ii) if a story contains three characters, then output the option with negative sentiment. Most stories in our dataset contained few characters. These rules are motivated by the intuition that a romantic story between two people can have a happy or sad ending depending on the context. However, a romantic story with three people is likely to describe a love triangle, and so not end well. Expectedly, these rules had low coverage (of about 60 stories), but a considerably high accuracy (70%) when active. Furthermore, a closer analysis revealed that most errors resulted from incorrect coreference resolutions (leading to incorrect count of characters). This indicates the utility of understanding semantics of social relationships for story comprehension and it could potentially be another aspect to consider while solving such tasks.

Sentiment Analysis: We now explore the insights obtained by modeling sentiments in stories. Mostafazadeh et al. (2016) presented two baselines for this task whose outputs were simply the *ending* whose sentiment agreed with (i) the complete story, or (ii) the *climax* (last sentence of the story). While their performances were close to random, our sentiment based features yield a much higher accuracy of 64.5% (see Table 2). This could possibly be attributed to our approach's ability to learn such rules from the data itself, rather than making hard assumptions. For instance, our language model of overall narrative sentiments in-

Context	Incorrect Ending	Correct Ending
My friends all love to go to the club to dance. They think it's a lot of fun and always invite. I finally decided to tag along last Saturday. I danced terribly and broke a friend's toe.	The next weekend, I was asked to please stay home.	My friends decided to keep inviting me as I am so much fun.
Johnny thought Anita was the girl for him, but he was wrong. He invited her out but she said she didn't feel well. Johnny decided to go to a club, just to drink and listen to music. At midnight, he looked back and saw Anita dancing with another guy.	Johnny did not ask Anita out again.	Johnny wanted to ask Anita out again.

Table 4: Examples of stories incorrectly predicted by our model.

icates that while happy stories mostly have happy endings (with a conditional probability of 74%), the reverse is not true. In particular, sad stories (with overall negative sentiments) end with a negative sentiment in only 52% of the cases. We made similar observations regarding sentimental conformity between endings and climaxes.

Our features' superior performance can also be attributed to their deeper understanding of not just overall sentiments but also their trajectories. Our language models indicate that stories that exhibit a positive sentiment in all three narrative segments (beginning, body, and climax) have very high chance of happy endings (83%). Similarly, stories with negative sentiments in the three segments also have a fair chance of having sad endings (60%). This is different from stories with an overall negative sentiment, in which case the sentiment may be exhibited in only certain narrative segments. The language models also identify a pattern of *hopeful* stories, in which the sentiment begins as negative but moves towards positive in the body and climax, resulting in mostly happy endings ($\sim 70\%$). This was not true for the reverse case: *pessimistic* stories with positive beginning but negative body (and/or climax) were equally likely to have positive or negative endings (52%). Supplementary material contains sample stories for each of the above observations.

4 Related Work

We now review previous work done in this field. Our work touches upon several research areas.

4.1 Story understanding:

Our work is most closely related to the field of narrative understanding. Apart from event-centric understanding of narrative plots (Lehnert, 1981;

McIntyre and Lapata, 2010; Goyal et al., 2010; Elsner, 2012; Finlayson, 2012), recent methods have focused on understanding narratives from the perspective of characters (Wilensky, 1978) mentioned in them. These methods study character personas (Bamman et al., 2013, 2014) or Proppian (Propp, 1968) roles (Valls-Vargas et al., 2014, 2015), inter-character relationships (Iyyer et al., 2016; Chaturvedi et al., 2016, 2017), and social networks of characters (Elson et al., 2010; Elson, 2012; Agarwal et al., 2013, 2014; Krishnan and Eisenstein, 2015; Srivastava et al., 2016).

4.2 Events-centered learning:

Our Entity-sequence component is closely related to semantic script learning. Script learning focuses on representing text using a prototypical sequences of events, their participants and causal relationships between them, called scripts (Schank and Abelson, 1977; Mooney and DeJong, 1985). Several statistical methods have been proposed to automatically learn scripts or scripts-like structures from unstructured text (Chambers and Jurafsky, 2008, 2009; Jans et al., 2012; Orr et al., 2014; Pichotta and Mooney, 2014). Such methods for script-learning also include Bayesian approaches (Bejan, 2008; Frermann et al., 2014), sequence alignment algorithms (Regneri et al., 2010) and neural networks (Modi and Titov, 2014; Granroth-Wilding and Clark, 2016; Pichotta and Mooney, 2016). There has also been work on representing events in a structured manner using schemas, which are learned probabilistically (Chambers, 2013; Cheung et al., 2013; Nguyen et al., 2015), using graphs (Balasubramanian et al., 2013) or neural approaches (Titov and Khoddam, 2015). Recently, Ferraro and Durme (2016) presented a unified Bayesian model for

scripts and frames.

4.3 Textual Coherence:

Our work is also related to the study of coherence in discourse. A significant amount of prior work is primarily based on the Centering Theory Framework (Grosz et al., 1995) and focus on entities and their syntactic roles (Karamanis, 2003; Karamanis et al., 2004; Lapata and Barzilay, 2005; Barzilay and Lapata, 2008; Elsner and Charniak, 2008). Other approaches measure coherence using topic drift within a domain (Barzilay and Lee, 2004; Fung and Ngai, 2006), co-occurrence of words (Lapata, 2003; Soricut and Marcu, 2006), syntactic patterns (Louis and Nenkova, 2012) and discourse relations (Pitler and Nenkova, 2008; Lin et al., 2011). The nature of the tasks addressed by these works (such as determining the correct arrangement order for a set of sentences) makes them focus on learning sequential order of the various discourse components (entities, ideas, etc.). Our goal, instead, is to choose between alternatives of discourse components themselves (and not just their order) to produce a consistent story.

5 Conclusion

Story comprehension is a complex Natural Language Understanding task involving linguistic intelligence as well as a semantic and social knowledge of the real world. This paper studies story comprehension from the perspective of learning what is likely to happen next in a story. We present a model that given a short story, predicts its correct ending. It incorporates three aspects of story-understanding, that are based on an analysis of the events, sentiments and topics described in the story. While this is the best-performing model till date on this task, our analysis indicates a need for even deeper analysis of human behavior and societal norms to further improve our understanding. This work emphasizes that there are multiple aspects to story understanding, which future research can build upon.

Acknowledgement

This work is partly supported by the IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM Cognitive Horizon Network; by the US Defense Advanced Research Projects Agency (DARPA) under contract FA8750-13-2-

0008; and by the Army Research Laboratory (ARL) under agreement W911NF-09-2-0053. The views expressed are those of the authors and do not reflect the official policy or position of IBM, the Department of Defense or the U.S. Government.

References

- Apoorv Agarwal, Sriramkumar Balasubramanian, Anup Kotalwar, Jiehan Zheng, and Owen Rambow. 2014. Frame semantic tree kernels for social network extraction from text. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*, pages 211–219.
- Apoorv Agarwal, Anup Kotalwar, and Owen Rambow. 2013. Automatic extraction of social networks from literary text: A case study on alice in wonderland. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013*, pages 1202–1208.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL 1998*, pages 86–90.
- Niranjan Balasubramanian, Stephen Soderland, Mausam, and Oren Etzioni. 2013. Generating coherent event schemas at scale. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1721–1731.
- Michael Bamberg. 2012. Narrative analysis. In D. L. Long A. T. Panter D. Rindskopf H. Cooper, P. M. Camic and K. Sher, editors, *APA handbook of research methods in psychology*, volume 2, pages 85–102. American Psychological Association.
- 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, ACL 2013*, 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, ACL 2014*, pages 370–379.
- Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.
- Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the Human Language Technology Confer-*

- ence of the North American Chapter of the Association for Computational Linguistics, *HLT-NAACL 2004*, pages 113–120.
- Cosmin Adrian Bejan. 2008. Unsupervised discovery of event scenarios from texts. In *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference*, pages 124–129.
- Nathanael Chambers. 2013. Event Schema Induction with a Probabilistic Entity-Driven Model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1797–1807.
- Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised Learning of Narrative Schemas and their Participants. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, ACL 2009*, pages 602–610.
- Nathanael Chambers and Daniel Jurafsky. 2008. Unsupervised Learning of Narrative Event Chains. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, ACL 2008*, pages 789–797.
- Eugene Charniak. 1972. Toward a model of children’s story comprehension. Technical report, Massachusetts Institute of Technology.
- Seymour Chatman. 1980. *Story and Discourse*. Cornell University Press.
- Snigdha Chaturvedi. 2016. *Structured Approaches for Exploring Interpersonal Relationships in Natural Language Text*. Ph.D. thesis, University of Maryland, College Park.
- Snigdha Chaturvedi, Mohit Iyyer, and Hal Daumé III. 2017. Unsupervised Learning of Evolving Relationships Between Literary Characters. In *Proceedings of the Thirty First AAAI Conference on Artificial Intelligence, AAAI’17*, pages 3159–3165.
- Snigdha Chaturvedi, Shashank Srivastava, Hal Daumé III, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA., AAAI’16*, pages 2704–2710.
- Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende. 2013. Probabilistic Frame Induction. In *Proceedings of the Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 837–846.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38.
- Micha Elsner. 2012. Character-based Kernels for Novelistic Plot Structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics EACL 2012*, pages 634–644.
- Micha Elsner and Eugene Charniak. 2008. Coreference-inspired coherence modeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, ACL 2008*, pages 41–44.
- David K. Elson. 2012. *Modeling Narrative Discourse*. Ph.D. thesis, Columbia University.
- David K. Elson, Nicholas Dames, and Kathleen McKown. 2010. Extracting Social Networks from Literary Fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010*, pages 138–147.
- Francis Ferraro and Benjamin Van Durme. 2016. A unified bayesian model of scripts, frames and language. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2601–2607.
- Mark Alan Finlayson. 2012. *Learning Narrative Structure from Annotated Folktales*. Ph.D. thesis, Massachusetts Institute of Technology.
- Lea Frermann, Ivan Titov, and Manfred Pinkal. 2014. A hierarchical bayesian model for unsupervised induction of script knowledge. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*, pages 49–57.
- Pascale Fung and Grace Ngai. 2006. One story, one flow: Hidden markov story models for multilingual multidocument summarization. *TSLP*, 3(2):1–16.
- Amit Goyal, Ellen Riloff, and Hal Daumé III. 2010. Automatically Producing Plot Unit Representations for Narrative Text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010*, pages 77–86.
- Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2727–2733.
- Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.
- Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry P. Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, CIKM 2013*, pages 2333–2338.

- Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan L. Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544.
- Bram Jans, Steven Bethard, Ivan Vulic, and Marie-Francine Moens. 2012. Skip n-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2012*, pages 336–344.
- Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. Topics in the Digital Humanities. University of Illinois Press.
- Sandra Jovchelovitch and Martin W. Bauer. 2000. Narrative interviewing. In Martin W. Bauer and G. Gaskell, editors, *Qualitative Researching With Text, Image and Sound : a Practical Handbook*, pages 57–74. SAGE, London, UK.
- Nikiforos Karamanis. 2003. *Entity Coherence for Descriptive Text Structuring*. Ph.D. thesis, Division of Informatics, University of Edinburgh.
- Nikiforos Karamanis, Massimo Poesio, Chris Mellish, and Jon Oberlander. 2004. Evaluating centering-based metrics of coherence. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, ACL 2004*, pages 391–398.
- Vinodh Krishnan and Jacob Eisenstein. 2015. “You’re Mr. Lebowski, I’m the Dude”: Inducing Address Term Formality in Signed Social Networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2015*, pages 1616–1626.
- Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, ACL 2003*, pages 545–552.
- Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, IJCAI 2005*, pages 1085–1090.
- Wendy G. Lehnert. 1981. Plot Units and Narrative Summarization. *Cognitive Science*, 5(4):293–331.
- Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006.
- Bing Liu, Mingqing Hu, and Junsheng Cheng. 2005. Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of the 14th international conference on World Wide Web, WWW 2005*, pages 342–351.
- Annie Louis and Ani Nenkova. 2012. A coherence model based on syntactic patterns. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012*, pages 1157–1168.
- Inderjeet Mani. 2012. *Computational Modeling of Narrative*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- Neil McIntyre and Mirella Lapata. 2010. Plot Induction and Evolutionary Search for Story Generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010*, pages 1562–1572.
- A. Mnih and G. Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th International Conference on Machine Learning, ICML 2007*, pages 641–648.
- Ashutosh Modi and Ivan Titov. 2014. Inducing neural models of script knowledge. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014*, pages 49–57.
- Raymond J. Mooney and Gerald DeJong. 1985. Learning schemata for natural language processing. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence, IJCAI 1985*, pages 681–687.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. 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 NAACL HLT 2016*, pages 839–849.
- Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.
- Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, and Romaric Besançon. 2015. Generative event schema induction with entity disambiguation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015*, pages 188–197.

- John Walker Orr, Prasad Tadepalli, Janardhan Rao Doppa, Xiaoli Fern, and Thomas G. Dietterich. 2014. Learning Scripts as Hidden Markov Models. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1565–1571.
- Haoruo Peng and Dan Roth. 2016. Two discourse driven language models for semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pages 290–300.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1532–1543.
- Karl Pichotta and Raymond J. Mooney. 2014. Statistical script learning with multi-argument events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*, pages 220–229.
- Karl Pichotta and Raymond J. Mooney. 2016. Learning statistical scripts with LSTM recurrent neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2800–2806.
- Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008*, pages 186–195.
- Vladimir Iakovlevich Propp. 1968. *Morphology of the Folktale*. University of Texas.
- Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dav Zimak. 2004. Semantic role labeling via integer linear programming inference. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING 2004*.
- Michaela Regneri, Alexander Koller, and Manfred Pinkal. 2010. Learning Script Knowledge with Web Experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010*, pages 979–988.
- Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures (Artificial Intelligence Series)*, 1 edition. Psychology Press.
- Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. Story cloze task: UW NLP system. In *Proceedings of L_S-D_{Sem}*, pages 52–55.
- Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, ACL 2006*, pages 1105–1112.
- Shashank Srivastava, Snigdha Chaturvedi, and Tom M. Mitchell. 2016. Inferring interpersonal relations in narrative summaries. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 2807–2813.
- Ivan Titov and Ehsan Khoddam. 2015. Unsupervised induction of semantic roles within a reconstruction-error minimization framework. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2015*, pages 1–10.
- Josep Valls-Vargas, Jichen Zhu, and Santiago Ontañón. 2014. Toward automatic role identification in unannotated folk tales. In *Proceedings of the Tenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2014*, pages 188–194.
- Josep Valls-Vargas, Jichen Zhu, and Santiago Ontañón. 2015. Narrative hermeneutic circle: Improving character role identification from natural language text via feedback loops. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, pages 2517–2523.
- Kurt Vonnegut. 1981. *Palm Sunday*. RosettaBooks LLC, New York.
- Robert Wilensky. 1978. *Understanding Goal-Based Stories*. Outstanding Dissertations in the Computer Sciences. Garland Publishing, New York.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, HLT/EMNLP 2005*, pages 347–354.