Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3500–3508
Marseille, 20-25 June 2022

A Corpus for Commonsense Inference in the Story Cloze Test

Bingsheng Yao, Ethan Joseph, Julian Lioanag, Mei Si
Rensselaer Polytechnic Institute
{yaob, josepe2, lioanj, sim}@rpi.edu

Abstract

The Story Cloze Test (SCT) is designed for training and evaluating machine learning algorithms for narrative understanding and inferences. The SOTA models can achieve over 90% accuracy on predicting the last sentence. However, it has been shown that high accuracy can be achieved by merely using surface-level features. We suspect these models may not *truly* understand the story. Based on the SCT dataset, we constructed a human-labeled and human-verified commonsense knowledge inference dataset. Given the first four sentences of a story, we asked crowd-source workers to choose from four types of narrative inference for deciding the ending sentence and which sentence contributes most to the inference. We accumulated data on 1871 stories, and three human workers labeled each story. Analysis of the intra-category and inter-category agreements show a high level of consensus. We present two new tasks for predicting the narrative inference categories and contributing sentences. Our results show that transformer-based models can reach SOTA performance on the original SCT task using transfer learning but don't perform well on these new and more challenging tasks.

Keywords: Narrative Reading Comprehension, Commonsense Reasoning, Story Cloze Test, Narrative Reasoning

1. Introduction

Narrative Reading Comprehension (RC) and commonsense reasoning are both prevalent Natural Language Processing (NLP) tasks. At the center of this intersection, story understanding is particularly challenging for machine learning algorithms and can even be hard for people (Charniak, 1972), (Oatley, 1994). Stories are more than sequences of words. Commonsense knowledge and reasoning are often needed to understand a story.

The Story Cloze Test (hereinafter, SCT-v1.0), published in 2016, is a popular dataset for training and evaluating machine story reasoning (Mostafazadeh et al., 2016). Each story is composed of five sentences and follows a character through a series of events to an ending event or situation. The first four sentences of each five-sentence story are provided as "context", and two alternative sentences are provided as the ending, labeled "correct" and "wrong". Each "context" and corresponding "correct" ending make up a complete five-sentence story. The goal is to predict the correct ending of the story. Table 1 shows two examples from the SCT-v1.0 dataset. Mostafazadeh et al. (2016) leveraged crowd-workers and carefully defined mechanisms to create the "wrong" endings and make sure they are entirely reasonable when read in isolation and not trivial.

Table 2 shows related work performance on the SCT-v1.0 test dataset. Mostafazadeh et al. (2016) provided results of several models at the time when SCT-v1.0 was published, the best of which achieved an accuracy of 59%. Latest works with transfer learning can reach accuracy comparable to human beings', which is 100%. Sun et al. (2018) achieved 88.3% accuracy by fine-tuning a GPT (Radford et al., 2018) model with training strategies such as back and forth reading, highlighting, and self-assessment. Cui et al. (2020) reported that the pre-trained BERT (Devlin et al., 2018) model can already achieve 89.2% and proposed Diff-Net along with BERT which can achieve 90.1% on SCT-v1.0. Li et al. (2019) also presented a transferable BERT model that can transfer language knowledge from a large-scale data corpus as well as various semantically related supervised tasks to achieve 91.8% accuracy.

Existing SOTA deep neural language models are approaching human performance. These results make it seems like the story understanding challenge has already been solved, at least for the SCT-v1.0. However, we notice that several earlier approaches performed relatively well on SCT-v1.0, even without understanding the meaning of the first four sentences. Most noticeably, Schwartz et al. (2017) leveraged stylistically linguistic features like word and character level n-grams for each ending sentence to build a classifier and achieved an accuracy of 75.2%, discarding the story content. This is not consistent with Mostafazadeh et al. (2016)' initiatives that are hoping to leverage narrative understanding as well as commonsense reasoning to tackle the SCT-v1.0 task. Further, Sharma et al. (2018) performed an extensive analysis on the SCT dataset and best performing models at that time, and pointed out that models leveraging human-authorship biases discovered in the SCT-v1.0 dataset can already achieve relatively high accuracy.

Based on these observations, we suspect that existing models do not have a deep understanding of the stories despite their performances. At least, the models should not be able to explain how and what knowledge is used to make their inferences. To test this hypothesis, we need labeling of the commonsense knowledge used in understanding stories in SCT-v1.0 which is missing in the original dataset.

Our work remedies this defect in the original SCT-v1.0 by leveraging human crowd-workers to identify reference sentences and the type of narrative reasoning they used for ending event inference in the SCT-v1.0 validation dataset, which contains 1871 data instances. The data is collected using Amazon mTurk. We provide four categories for the reasoning of story ending, which are behavior-based, objective-based, emotional-based, and goal-driven. The first four sentences are given to the crowd-workers, and they are asked to infer the fifth ending sentence. The crowd-workers are asked to think about and categorize how they made the decisions based on the four categories we provided. We require each story's data to be labeled by three workers to minimize individual biases. We hypothesize that even if a model can reach human-level performance at the original SCT-v1.0 task, it may not correctly identify the type of narrative reasoning required to conclude. To test this hypothesis, we create two tasks for this new data: commonsense inference category prediction and commonsense prime determinant sentence prediction. Only when a model can correctly perform these tasks in addition to predicting the endings of the stories, we will have confidence that the model possesses a deep understanding of the story and can make judgments in narrative understanding as human beings do.

This paper makes the following contributions:

- We provide a new crowd-sourced dataset that augments the original story cloze dataset with the type of reasoning and evidence sentences for choosing the right ending.
- We propose two new tasks based on this new dataset – the prediction of inference category and prime determinant sentence.
- We demonstrate that a transformer-based model can reach human-level performance on the original SCT-v1.0 task, yet still has large room to improve on these new tasks. Therefore, we believe this new dataset and new tasks can help AI models to gain a deeper understanding of the story.

2. Background

2.1. Reading Comprehension and Commonsense Reasoning Tasks

Reading Comprehension (RC) with commonsense reasoning is a task to read and comprehend given text chunks and articles and complete a variety of tasks based on the text corpus by leveraging commonsense knowledge embedded into the text, such as cloze-style RC, open-domain RC. The early (Chambers and Jurafsky, 2009) work on the narrative cloze test did not rely on the order of the sentences. Three years later, Jans et al. (2012) redefined the test to be an ordered sequence of events. A few frameworks for evaluating language RC have been developed since then. For example, MCTest (Richardson et al., 2013) involves story comprehension, but they are mostly targeting fictional and children's stories. Mostafazadeh et al. (2016) presented ROCStories, a collection of 50k high-quality five-sentence commonsense stories about everyday life, and the Story Cloze Test (SCT-v1.0), to serve as a proper evaluation framework for Commonsense RC correspondingly.

Despite the SCT-v1.0 dataset intended to require machines with reading comprehension capabilities for the inference, existing neural models do not necessarily need to understand the story for performing the task well. For example, a linear model Cogcomp (Khashabi et al., 2018) leveraged sentiment trajectory, topical consistency, and event sequences and performed with 74.4% accuracy. MSAP (Schwartz et al., 2017) used stylistically linguistic features like word and character level n-grams for each ending sentence to build a classifier and achieved an accuracy of 75.2% on SCT-v1.0, which fully discarded the text context.

Sharma et al. (2018) performed an extensive analysis on the SCT-v1.0 dataset and the best performing models at that time. Sharma et al. (2018) pointed out that models leveraging human-authorship biases discovered in the SCT-v1.0 dataset can already achieve relatively high accuracy. Sharma et al. (2018) proposed an updated SCT-v1.0 dataset trying to overcome some of the biases they have discovered, hereinafter SCT-v1.5. However, the latest works with pre-trained or transfer learning encoder-decoder structured deep neural models can achieve over 80% accuracy on both SCT-v1.0 and SCT-v1.5. Sun et al. (2018) achieved 88.3% accuracy by fine-tuning a GPT model, which is a generative pretrained transformer (Radford et al., 2018), with training strategies such as back and forth reading, highlighting, and self-assessment. Cui et al. (2020) reported a pretrained BERT model, which is a Bidirectional Encoder Representation from Transformers (Devlin et al., 2018), can achieve 89.2% and Cui et al. (2020) proposed DiffNet alone with BERT can achieve 90.1% on SCT-v1.0, 82.0% accuracy on SCT-v1.5 correspondingly. Li et al. (2019) also presented a transferable BERT model that can transfer language knowledge from large-scale data corpus as well as various semantically related supervised tasks, to achieve 91.8% accuracy on SCT-v1.0 and 90.3% on SCT-v1.5.

2.2. Narrative Comprehension

When designing the categories for how story endings are reasoned from the previous sentences, we draw inspiration from existing narrative comprehension theories. Studied through the formalisms of cognitive psychology, the structuralist view of narrative led to models of narrative comprehension that focused on how people understand and represent information. These models mapped to models of knowledge representation, such as the adoption of Minsky's frames (Minsky, 1974) by Kintsch, van Dijk, and Teun to account for the comprehension of semantics in narrative discourse (Van Dijk, 1977; Kintsch and Van Dijk, 1978), and Schank & Abelson's scripts (Schank and Abelson, 1975; Abelson, 1981), schema-like structures for typified action sequences which were studied in relation to how people read and recall narratives (Bower et al., 1979).

First Four Sentences	Correct Ending	Incorrect Ending
Mary makes candles. She enjoys this as a hobby. Mary took one to her friend as a gift. Everyone loved the candles and asked to buy one.	Mary was very happy.	Mary felt unappreciated.
Sam loved his old belt. He matched it with everything. Unfortunately he gained too much weight. It became too small.	Sam went on a diet.	Sam was happy.

Table 1: Example of stories and ending alternatives in the SCT-v1.0 dataset. Given the first four sentences of a story, the original task is to predict the correct ending sentence.

Authors	Prediction accuracy
Mostafazadeh et al. (2016)	59%
Khashabi et al. (2018)	74.4%
Schwartz et al. (2017)	75.2%
Sun et al. (2018)	88.3%
Cui et al. (2020)	90.1%
Li et al. (2019)	91.8%

Table 2: Prediction accuracy of existing works on the original SCT-v1.0 dataset.

Tannen describes the various models used to study narrative comprehension as “structures of expectation”, a term borrowed from Ross in his study of semantic understanding in discourse (Ross, 1975; Tannen, 1977). By structures of expectation, Tannen refers to how the various models represent the ways in which people organize the information they have previously encountered in order to parse new information (Tannen, 1993), an interpretation in line with Bartlett’s initial formulation of schema (Bartlett and Bartlett, 1995) and Weick’s formulation of sensemaking (Weick et al., 2005).

Gernsbacher & Givon categorize grammatical markers of coherence in human-written text into spatial, temporal, and referential continuity (Givón, 1992; Gernsbacher and Givón, 1995). In a coding scheme for personal narratives designed for use in clinical psychology, researchers (Reese et al., 2011) categorize elements of narrative coherence as spatial orientation, temporal orientation, and a larger element they call topical consistency. Topical consistency encompasses causal links, personal emotional and motivational evaluations, and connections to previous stories or events in a person’s life (Reese et al., 2011). Rideout studied personal narratives delivered in court cases and observed causal and character consistency as major elements of narrative coherence, including consistency of character motivations and causal links between events (Rideout, 2013). Consistency of character motivations relates more broadly to general character consistency and believability, which have been pointed out as major components of people’s feelings of coherence in narratives (Riedl and Young, 2010; Rapp et al., 2001). Based on our observation of the five-sentences stories in SCT-v1.0, we found topical consistency as described in (Reese et al., 2011) being the most relevant narrative coherence marker. We further divide the inference process into four types: behavior-based, objective-based, emotional-based, and goal-driven, which are introduced in the next section.

3. Human Labeling Task

Our goal is to extend the validation split of the SCT-v1.0 dataset by asking human annotators to not only explicitly categorize the type of narrative coherence reasoning that they used for the ending sentence inference but also label the key sentences that inference is based on. We will then require consensus from multiple human annotators.

Categories of Narrative Coherence Reasoning: We propose a four-category classification for the commonsense knowledge used by human annotators while inferring the ending sentence: Behavior-based means the event is an action by the character’s own initiative; Objective-based corresponds to an objective or external environment that causes the character to react; Emotional-based represents descriptions about a character’s feelings or other emotional states; Goal-driven explicitly describes character’s goals and targets. Examples are shown in Table 3.

Procedure: We first perform a pilot study with college students on 100 randomly selected stories from SCT-v1.0 dataset. This study aims to make sure that the instructions are clear and the four categories we designed are meaningful. Then, we collect the rest of the data by leveraging crowd workers from Amazon mTurk. During the mTurk collecting process, we first present the dataset and the classification task with examples to the crowd-workers. Stories are randomly assigned to the workers while making sure that each story is labeled by three workers. We set a limit that any worker can only annotate up to 50 stories so as to prevent the data from being heavily influenced by any individual.

Interface: Figure 1 shows the interface of the annotation task on mTurk. The workers can indicate that the narrative inference belongs to one or multiple categories by selecting the key reference sentences under those categories. Each sentence is only allowed to be selected for one category, but the workers are allowed to choose multiple sentences for one category. Further, we asked the workers to self-report their confidence levels regarding the selections they made for each category. We require the crowd-workers to be qualified mTurk Masters, and the annotation task for each story needs to

Category	First 4 Sentences of the Story	Ending Sentence
Behavior-based	The building exploded. Five people were inside it. Three died and two lived. ***Mosh rushed out with his little sister in his arms.***	He was relieved to be alive.
Objective-based	My brother got a ticket. He never went to court. He eventually got pulled over again. ***They arrested him for failure to appear.***	He went to jail.
Emotional-based	Gina went to her 6th hour class. It was gym. Gina hated gym, and all things related. ***She was horrified when the teacher made them run the first day.***	Gina cried at the end.
Goal-driven	My girlfriend collects stamps. Not the kind for mail but the kind to stamp things with. She has different cute ones. ***For her birthday I want to buy her more.***	I bought her a rare stamp from an online seller.

Table 3: Example of each category for the Commonsense Inference in SCT-v1.0. We highlight the determinant sentence from the first four sentences and the ending sentence.

be finished within five minutes.

4. Quantitative Analysis

We perform statistical and quantitative analysis of human consensus on the combined pilot study and mTurk study data.

4.1. Inter-category Consensus Analysis

Table 4 shows three typical examples of human consensus. Each row represents the annotation of a single worker. The number before the slash is the index of the sentence showing which sentence(s) are classified into each category by the mTurk workers. The value after the slash in each cell is the normalized relevance value given by the workers. A cell without an index means that none of the sentences is selected for this category. As we can observe from the table, there is a high consensus over the prime determinant sentence or the most important category for commonsense inferences. For the first example, three workers share exactly the same sentence classification, except that they weigh differently on the relevance value. In this case, we determine the sentence with the highest sum of relevance value, which is the first sentence, to be the prime determinant sentence and the corresponding category. Therefore, the behavior-based category is determined to be the most important category for commonsense inference. In the second example, we observe that all three workers have a consensus over selecting and classifying the first sentence to be behavior-based and the fourth sentence to be the emotional-based category. The fourth sentence becomes the prime determinant sentence because the sum of relevance value is higher than that of the first sentence. Though the sentence classifications in the third example vary a lot among different workers, we can easily observe that they all agree on the fourth sentence to be the objective-based category and the fourth sentence is the only sentence being selected by all workers. This tie-breaking algorithm is being leveraged to prepare subset data for the commonsense inference category prediction task, and we show details of this algorithm in Algorithm 1.

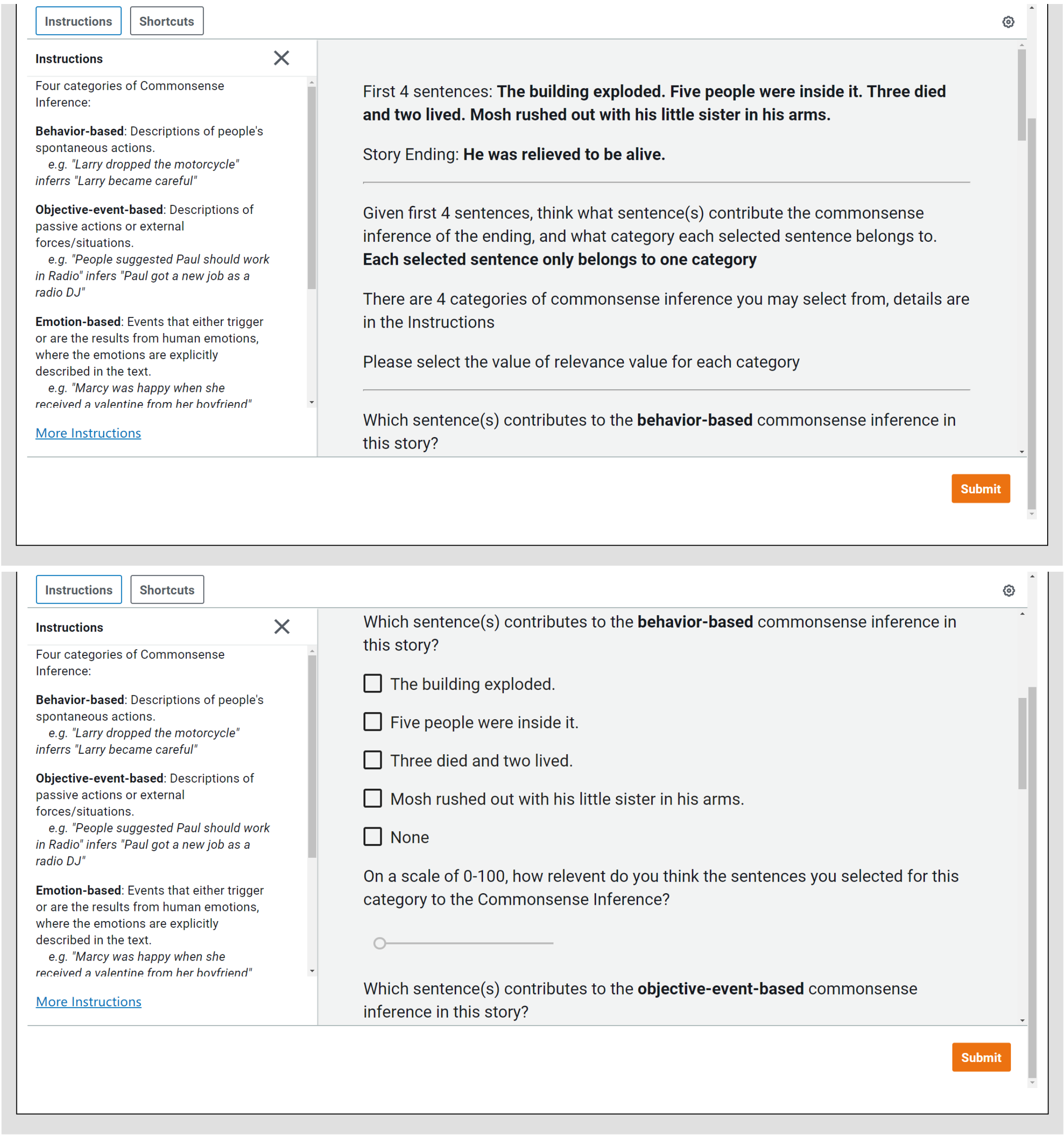
4.2. Intra-category Consensus Analysis

We further analyze the consensus of commonsense inference within each category. For each category, we first count how many stories that two or more workers annotate at least one sentence into that category. To further analyze to what extent humans agree on classifying the same sentence into the same category, we calculate how many stories have such labeling that the same sentence is selected by two or more workers and classified into the same category. Results are shown in Figure 2. For instance, there are 1869 stories have two or more workers classify at least one sentence into the behavior-based category, and 1502 stories have two or more workers classify the same sentence into the behavior-based category. We observe that workers are highly likely to classify the same sentence into the same category if they are certain about the category of the commonsense reasoning being involved in the inference. More specifically, there are 74.34% to 80.36% cases where two or more workers classify the same sentence into the same category compared with all the cases where this category is being selected.

5. Two Commonsense Inference Tasks

The quantitative consensus analysis indicates a relatively high level of agreement among the mTurk workers. Therefore, we propose two new tasks: commonsense inference category prediction and commonsense prime determinant sentence prediction. Our proposed tasks focus on how the machine learning models can mimic humans' use of narrative commonsense knowledge to make inferences. Specifically, we ask the models to predict which type of narrative inference is most important for picking the correct story endings and which sentence(s) are the most critical determinants of the inference.

To prepare data for the commonsense inference category prediction task, we select a subset of labeled story dataset with clear human label consensus on a unique top category classification. We determine whether there exists clear consensus via a 3-step tie-breaking algorithm shown in Algorithm 1. First, we initialize (line 6)



Instructions

Four categories of Commonsense Inference:

Behavior-based: Descriptions of people's spontaneous actions.
e.g. "Larry dropped the motorcycle" inferrs "Larry became careful"

Objective-event-based: Descriptions of passive actions or external forces/situations.
e.g. "People suggested Paul should work in Radio" infers "Paul got a new job as a radio DJ"

Emotion-based: Events that either trigger or are the results from human emotions, where the emotions are explicitly described in the text.
e.g. "Marcy was happy when she received a valentine from her boyfriend"

More Instructions

First 4 sentences: **The building exploded. Five people were inside it. Three died and two lived. Mosh rushed out with his little sister in his arms.**

Story Ending: **He was relieved to be alive.**

Given first 4 sentences, think what sentence(s) contribute the commonsense inference of the ending, and what category each selected sentence belongs to. **Each selected sentence only belongs to one category**

There are 4 categories of commonsense inference you may select from, details are in the Instructions

Please select the value of relevance value for each category

Which sentence(s) contributes to the **behavior-based** commonsense inference in this story?

Instructions

Four categories of Commonsense Inference:

Behavior-based: Descriptions of people's spontaneous actions.
e.g. "Larry dropped the motorcycle" inferrs "Larry became careful"

Objective-event-based: Descriptions of passive actions or external forces/situations.
e.g. "People suggested Paul should work in Radio" infers "Paul got a new job as a radio DJ"

Emotion-based: Events that either trigger or are the results from human emotions, where the emotions are explicitly described in the text.
e.g. "Marcy was happy when she received a valentine from her boyfriend"

More Instructions

Which sentence(s) contributes to the **behavior-based** commonsense inference in this story?

☐ The building exploded.

☐ Five people were inside it.

☐ Three died and two lived.

☐ Mosh rushed out with his little sister in his arms.

☐ None

On a scale of 0-100, how relevent do you think the sentences you selected for this category to the Commonsense Inference?

Which sentence(s) contributes to the **objective-event-based** commonsense inference in this story?

Figure 1: Annotation interface for the mTurk workers. We also provide examples of each commonsense inference category in instructions.

three counters that count the number of times each category is selected, the sum of relevance values for each category, and the total number of sentences that fall into each category. For each story in the dataset, we loop through the labeling for each category created by each of the three workers and increment the counter correspondingly. Then the 3-step tie-breaking algorithm contains three rules which we apply in sequence: 1) return the category if there exists a unique category being selected the most number of times among three workers (line 21); 2) otherwise, compare the sum of relevance value among the categories with maximum selection counts, return if there is a unique category with the biggest sum (line 23); and 3), choose the category with the most amount of sentences among the top-category candidates of the previous steps if there is still a tie (line 25).

A subset of 1619 stories is obtained with the tie-breaking algorithm above for the commonsense inference category prediction task. We further split this subset following an 80/10/10 split to get 1295 samples for the training set, 162 for validation, and 162 for testing. We notice that in this subset, three categories – behavior-based, objective-based, and goal-driven – have relatively even distributions (19.6%, 19.7%, 15.4% respectively), and the emotional-based category has a much higher

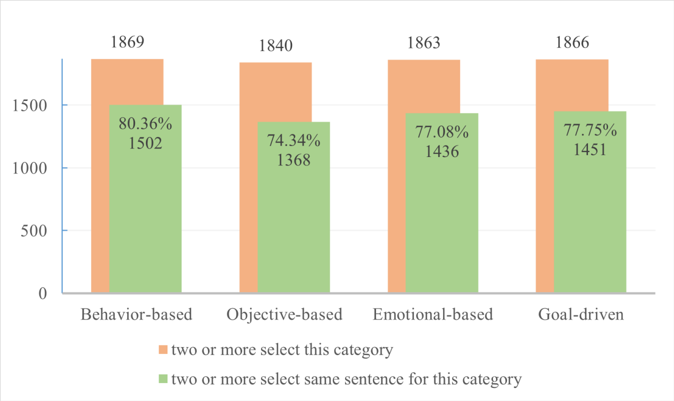


Figure 2: Human consensus on sentence selection and categorization of Commonsense Inference for SCT-v1.0 Val with 1871 stories in total. Workers are likely to classify the same sentence into the same category if they are certain about the inference category.

Story content	Behavior-based	Objective-based	Emotional-based	Goal-driven
1. Matt was out for a walk with his mom. 2. Suddenly he saw something shiny on the ground. 3. He bent to pick it up. 4. It was a gold ring! 5. Matt was excited to be so lucky.	**1. / 0.26**	3. / 0.26	4. / 0.26	2. / 0.2
	1. / 0.23	3. / 0.29	4. / 0.23	2. / 0.23
	1. / 0.28	3. / 0.21	4. / 0.21	2. / 0.28
1. I cleaned my wedding ring. 2. I first set it in the cleaner to soak. 3. I then wiped it with the brush to make sure I got everything off. 4. Then I put it back in the cleaner for a bit. 5. It came out sparkling.	1. / 0.18	3. / 0.36	**4. / 0.27**	2. / 0.18
	1. / 0.27	/	**4. / 0.36**	3. / 0.36
	1. / 0.25	3. / 0.25	**4. / 0.25**	2. / 0.25
1. Bobby was having a birthday party. 2. His friends bought him a huge cake. 3. A woman jumped out of the cake. 4. Bobby was surprised and amused. 5. Bobby had a good party.	3. / 0.4	**4. / 0.4**	/	2. / 0.2
	3. / 0.25	**4. / 0.25**	2. / 0.25	1. / 0.25
	1. / 0.27	**4. / 0.18**	3. / 0.27	2. / 0.27

Table 4: Example commonsense inference labelings that form different consensus on SCT-v1.0 validation dataset by crowd workers.

count of roughly 45.1% of the data. We assume two potential reasons to explain the unbalance and are worth further analysis: 1. The original dataset may not be evenly distributed, and 2. People may be biased to choose the emotional-based category as long there are emotional descriptions in the story instead of thinking about the reasoning determinant.

The second task, the commonsense prime determinant sentence prediction task, is more challenging because people may leverage different sentences as the determinant for Commonsense Inference, and both could be reasonable. We create another subset for this task by counting how many times each sentence is selected by three workers and only consider stories that have human consensus on top one or two determinant sentence(s). The subset for this task contains 1074 stories. We follow the same split as the category task to obtain 859 samples for training, 108 for validation, and 107 for testing.

Our intention for designing these two tasks is to guide machine learning models to mimic the human process of reasoning with narrative commonsense knowledge. As mentioned before, we don't believe the SOTA models for the SCT-v1.0 can automatically perform well on these tasks. To establish a baseline, we test fine-tuning transformer-based language models on these two datasets, as these models have reached SOTA performance on the original SCT-v1.0 task. More specifically, we test the performance of these three models: BERT (Devlin et al., 2018), RoBERTA (Liu et al., 2019), and DeBERTA (He et al., 2020). We expect that they will all perform well on SCT-v1.0 but fail the two new tasks we proposed.

6. Baseline Model Performance

We evaluated the performance of three SOTA transformer models, BERT, RoBERTa, and DeBERTa, to serve as baselines on the proposed tasks. Additionally, we evaluated the models' performance on SCT-v1.0 for comparison with previous SOTA. Each model was fine-tuned on an NVIDIA Tesla T4 for up to an hour depending on the task. We fine-tuned each model on the training sets and report performance metrics for the fine-tuned models on the test sets in Table 5.

6.1. Training Setup

In each task, we fine-tune a transformer language model pretrained with a masked language modeling objective. For the SCT-v1.0 task, the 4 sentences of a story and the two endings (wrong and correct) are concatenated and then tokenized with WordPiece (Wu et al., 2016) for

Test	SCT-v1.0	Inference Category			Prime Determinant Sentence		
Model	Accuracy	Accuracy	Balanced Acc.	Macro F1	Accuracy	Hamming Loss	Macro F1
BERT-Large	78.2%	37.7%	24.8%	36.1%	20.4%	**0.354**	34.3%
RoBERTa-Large	93.1%	37.0%	25.9%	35.9%	29.6%	0.303	**45.8%**
DeBERTa-Large	93.1%	**46.3%**	**30.2%**	**43.1%**	**31.5%**	0.313	41.1%

Table 5: Baseline Comparisons of BERT, RoBERTa, and DeBERTa on Test Sets

Algorithm 1 Tie-breaking to find the top-category for each story

```
Input: human labellings on one story $L$
Output: top-category $c$ selected for this story, None otherwise
Data: Complete human labellings on $SCT-v1.0Val$
▷Each story is labeled by 3 human workers
assert $len(L) == 3$
$cnt\_category = cnt\_rv = cnt\_sent = [0] * 4$
for $l_i$ in $L$ do                ▷ $C$ : list of 4 categories
    for $c_i$ in $C$ do
        if $l_i[c_i]$ is not None then
            $cnt\_category[c_i]$ += 1
            $cnt\_rv[c_i]$ += relevance value $rv_{c_i}$
            $cnt\_sent[c_i]$ += sentence count
        end if
    end for
end for
$max_{cnt} \leftarrow \max(cnt\_category)$
$max_{rv} \leftarrow \max(cnt\_rv)$
$max_{sent} \leftarrow \max(cnt\_sent)$
▷3-step tie-breaking strategy
if unique $cnt\_category.count(max_{cnt})$ then
    return $C[cnt\_category.index(max_{cnt})]$
else if unique $cnt\_rv.count(max_{rv})$ and this category has $max_{cnt}$ then
    return $C[cnt\_rv.index(max_{rv})]$
else if unique $cnt\_sent.count(max_{sent})$ and this category has both $max_{cnt}$ and $max_{rv}$ then
    return $C[cnt\_sent.index(max_{sent})]$
else
    return None
end if
```

BERT and DeBERTa, and byte-pair encoding (Sennrich et al., 2016) for RoBERTa. For the proposed inference category and prime determinant sentence tasks, the 5 sentences of a story are concatenated and then tokenized the same way as in SCT-v1.0.

For the SCT-v1.0 task (binary classification), we fine-tune the models to minimize a cross-entropy loss between their predicted category and the true category label:

$$CE = -\sum_{c}^{2} y_{o,c} \log(p_{o,c})$$

where y is a binary indicator if class label c is the correct classification for observation o, and p is the predicted probability observation o is of class c. The model outputs a softmax probability over the two classes (ending 1 vs. ending 2). We then take the higher probability class as the prediction. To serve as a comparison between previous works on SCT-v1.0, we evaluated the models on accuracy as it is the most commonly reported metric on this dataset (See Table 5).

For the commonsense inference category prediction task (multiclass classification), we fine-tune the models to minimize a weighted cross-entropy loss between their predicted category and the true category label:

$$CE = -\sum_{c}^{4} w_c (y_{o,c} \log(p_{o,c}))$$

where w_c is the weight for class label c, y is a binary indicator if class label c is the correct classification for observation o, and p is the predicted probability observation o is of class c. The model outputs a softmaxed probability over all class labels (behavior-based, objective-based, emotional-based, and goal-driven). For this task, the class with the highest probability serves as the prediction. We then evaluate the models on accuracy, balanced accuracy, and macro F1 score (See Table 5).

For the commonsense prime determinant sentence task (multilabel classification), we fine-tune the models to minimize a binary cross-entropy loss between their predicted sentences and the true sentences label for each possible sentence prediction:

$$CE = -(y \log(p) + (1 - y) \log(1 - p))$$

where y is a binary indicator if class label is the correct classification, and p is the predicted probability. The model outputs a probability between 0 and 1 for each sentence separately (with 4 total sentences). During inference, we set a discrimination threshold of 0.5 and sentences with a probability $>$ 0.5 are the prediction. We then evaluate the models on multiple evaluation metrics, including accuracy, hamming loss, and macro F1 score (See Table 5).

6.2. Results and Discussion

In Table 5, we can see that for SCT-v1.0, all three models performed reasonably well. RoBERTa and DeBERTa have a much higher accuracy rate than the standard BERT model. The differences between RoBERTa and

DeBERTa are further explored in the Inference Category and Prime Determinant Sentence results, where DeBERTa outshines RoBERTa on accuracy, balanced accuracy and Macro F1 of the Inference Category test as well as accuracy and hamming loss of Prime Determinant Sentence test, where RoBERTa only has an advantage on the Macro F1 of Prime Determinant Sentence test over DeBERTa. It is clearly evident through baseline comparisons that DeBERTa is the best performing model for all three tests.

Furthermore, we can see that accuracy on the SCT-v1.0 has no stake in the performance of all models on the Inference Category and Prime Determinant Sentence tests. Each model performs significantly worse in the two new tasks than in SCT-v1.0.

7. Conclusion and Future Work

A key challenge in narrative understanding is the lack of high-quality data that labels the narrative strategies and phenomena in stories. We propose two new narrative understanding tasks and collect rich human-labeled and human-verified datasets for narrative commonsense inference and prime determinant sentences identification based on the SCT-v1.0 dataset.

We demonstrate that although the SOTA neural models have been proven to perform very well on the original SCT-v1.0 binary classification tasks of choosing the correct ending sentence, our two proposed tasks remain very challenging to the models. This is because both tasks require the models to understand and leverage the narrative commonsense knowledge underlying. We believe these two new datasets and the proposed tasks will help train neural models to better understand and utilize narrative commonsense knowledge. This project lays the groundwork for creating explainable AI models for narrative understanding.

For future work, we will explore improving the models' performances by supplementing story structure and character modeling related to commonsense knowledge. The stories in SCT-v1.0 are short and relatively simple. We are also interested in collecting additional human-annotated data for different types of stories. Finally, we want to explore downstream tasks that can benefit from having the model learn our proposed two tasks first. For example, we hope to find out whether these two tasks can help a model answer other types of narrative commonsense questions or generate better stories.

8. References

Abelson, R. P. (1981). Psychological status of the script concept. *American psychologist*, 36(7):715.

Bartlett, F. C. and Bartlett, F. C. (1995). *Remembering: A study in experimental and social psychology*, volume 14. Cambridge University Press.

Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cognitive psychology*, 11(2):177–220.

Chambers, N. and Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610.

Charniak, E. (1972). *Toward a model of children's story comprehension.* Ph.D. thesis, Massachusetts Institute of Technology.

Cui, Y., Che, W., Zhang, W.-N., Liu, T., Wang, S., and Hu, G. (2020). Discriminative sentence modeling for story ending prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7602–7609.

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

Gernsbacher, M. A. and Givón, T. (1995). *Coherence in spontaneous text*, volume 31. John Benjamins Publishing.

Givón, T. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics*, 30(1):5–56.

He, P., Liu, X., Gao, J., and Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654.*

Jans, B., Bethard, S., Vulić, I., and Moens, M. F. (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*, pages 336–344.

Khashabi, D., Sammons, M., Zhou, B., Redman, T., Christodoulopoulos, C., Srikumar, V., Rizzolo, N., Ratinov, L., Luo, G., Do, Q., et al. (2018). Cogcompnlp: Your swiss army knife for nlp. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Kintsch, W. and Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review*, 85(5):363.

Li, Z., Ding, X., and Liu, T. (2019). Story ending prediction by transferable bert. *arXiv preprint arXiv:1905.07504.*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

Minsky, M. (1974). A framework for representing knowledge.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., and Allen, J. (2016). A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696.*

Oatley, K. (1994). The creative process: A computer model of storytelling and creativity scott r. turner.

Radford, A., Narasimhan, K., Salimans, T., and

Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf.*

Rapp, D. N., Gerrig, R. J., and Prentice, D. A. (2001). Readers' trait-based models of characters in narrative comprehension. *Journal of Memory and Language*, 45(4):737–750.

Reese, E., Haden, C. A., Baker-Ward, L., Bauer, P., Fivush, R., and Ornstein, P. A. (2011). Coherence of personal narratives across the lifespan: A multidimensional model and coding method. *Journal of Cognition and Development*, 12(4):424–462.

Richardson, M., Burges, C. J., and Renshaw, E. (2013). Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203.

Rideout, J. C. (2013). A twice-told tale: Plausibility and narrative coherence in judicial storytelling. *Legal Comm. & Rhetoric: JAWLD*, 10:67.

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

Ross, R. N. (1975). Ellipsis and the structure of expectation. *San Jose State Occasional Papers in Linguistics*, 1(18):3–9.

Schank, R. C. and Abelson, R. P. (1975). Scripts, plans, and knowledge. In *IJCAI*, pages 151–157.

Schwartz, R., Sap, M., Konstas, I., Zilles, L., Choi, Y., and Smith, N. A. (2017). The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. *arXiv preprint arXiv:1702.01841*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sharma, R., Allen, J., Bakhshandeh, O., and Mostafazadeh, N. (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.

Sun, K., Yu, D., Yu, D., and Cardie, C. (2018). Improving machine reading comprehension with general reading strategies. *arXiv preprint arXiv:1810.13441*.

Tannen, D. (1977). Well what did you expect? In *Annual Meeting of the Berkeley Linguistics Society*, volume 3, pages 506–515.

Tannen, D. (1993). What's in a frame? surface evidence for underlying expectations. *Framing in discourse*, 14:56.

Van Dijk, T. A. (1977). Semantic macro-structures and knowledge frames in discourse comprehension. *Cognitive processes in comprehension*, 332.

Weick, K. E., Sutcliffe, K. M., and Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization science*, 16(4):409–421.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.