# An Encoder-decoder Approach to Predicting Causal Relations in Stories

**Melissa Roemmele and Andrew S. Gordon**
Institute for Creative Technologies, University of Southern California
roemmele@ict.usc.edu, gordon@ict.usc.edu

## Abstract

We address the task of predicting causally related events in stories according to a standard evaluation framework, the Choice of Plausible Alternatives (COPA). We present a neural encoder-decoder model that learns to predict relations between adjacent sequences in stories as a means of modeling causality. We explore this approach using different methods for extracting and representing sequence pairs as well as different model architectures. We also compare the impact of different training datasets on our model. In particular, we demonstrate the usefulness of a corpus not previously applied to COPA, the ROCStories corpus. While not state-of-the-art, our results establish a new reference point for systems evaluated on COPA, and one that is particularly informative for future neural-based approaches.

## 1 Introduction

Automated story understanding is a long-pursued task in AI research (Dehn, 1981; Lebowitz, 1985; Meehan, 1977). It has been examined as a commonsense reasoning task, by which systems make inferences about events that prototypically occur in common experiences (e.g. going to a restaurant) (Schank and Abelson, 1977). Early work often failed to scale beyond narrow domains of stories due to the difficulty of automatically inducing story knowledge. The shift to data-driven AI established new opportunities to acquire this knowledge automatically from story corpora. The field of NLP now recognizes that the type of commonsense reasoning used to predict what happens next in a story, for example, is as important for natural language understanding systems as linguistic knowledge itself.

A barrier to this research has been the lack of standard evaluation schemes for benchmarking progress. The Story Cloze Test (Mostafazadeh et al., 2016) was recently developed to address this, with a focus on predicting events that are temporally and causally related within common real-world scenarios. The Story Cloze Test involves selecting which of two given sentences best completes a particular story. Related to this is the Choice of Plausible Alternatives (COPA) task (Roemmele et al., 2011), which uses the same binary-choice format to elicit a prediction for either the cause or effect of a given story event. While the Story Cloze Test involves predicting the ending of a story, COPA items focus specifically on commonsense knowledge related to identifying causal relations between sequences.

The competitive approaches to narrative prediction evaluated by the Story Cloze Test largely involve neural networks trained to distinguish between correct and incorrect endings of stories (Cai et al., 2017, e.g.). A neural network approach has yet to be applied to the related COPA task. In the current paper, we initiate this investigation into these models for COPA. In particular, we evaluate an encoder-decoder model that predicts the probability that a particular sequence follows another in a story. Our experiments explore a few different variables for configuring this approach. First, we examine how to extract temporally related sequence pairs provided as input to the model. Second, we vary the use of feed-forward versus recurrent layers within the model. Third, we assess different vector-based representations of the sequence pairs. Finally, we compare our model using different narrative corpora for training, including the ROCStories corpus which was developed in conjunction with the Story Cloze Test. Our results are presented in comparison to existing systems applied to COPA, which involve lexical co-occurrence statistics gathered from web corpora. Our best-performing model achieves an accuracy of 66.2% on the COPA test set, which falls short of

the current state-of-the-art of 71.2% (Sasaki et al., 2017). Interestingly, this best result utilizes the ROCStories for training, which is only a small fraction of the size of the datasets used in existing approaches. Applying our model to these larger datasets actually yields significantly worse performance, suggesting that the model is sensitive to the density of commonsense knowledge contained in its training set. We conclude that this density is far more influential to COPA performance than just data quantity, and further success on the task will depend on methods for isolating implicit commonsense knowledge in text.

## 2 Choice of Plausible Alternatives

The Choice of Plausible Alternatives (COPA) is composed of 1,000 items, where each item contains three sentences, a *premise* and two *alternatives*, as well as a prompt specifying the relation between them. The items are divided equally into development and test sets of 500 items each. The goal is to select which alternative conveys the more plausible cause (or effect, based on the prompt) of the premise sentence. Half of the prompts elicit the more plausible effect of the premise event, while the other half ask for the more plausible cause of the premise.

1. **Premise:** The homeowners disliked their nosy neighbors. *What happened as a result?*
   **Alternative 1:*** They built a fence around their property.
   **Alternative 2:** They hosted a barbecue in their backyard.

2. **Premise:** The man fell unconscious. *What was the cause of this?*
   **Alternative 1:*** The assailant struck the man in the head.
   **Alternative 2:** The assailant took the man's wallet.

Above are examples of COPA items, where the designated correct alternative for each is starred. In a given item, both alternatives refer to events that could be found within the same story, but the correct one conveys a more coherent causal relation. All sentences consist of a single clause with a past tense verb. COPA items were written by a single author and then validated by other annotators to ensure human accuracy approximated 100%. See Roemmele et al. (2011) for further details about the authoring and validation process.

## 3 Existing Approaches

Roemmele et al. (2011) presented a baseline approach to COPA that focused on lexical co-occurrence statistics gathered from story corpora. The general idea is that a causal relation between two story events can be modeled by the proximity of the words that express the events. This approach uses the Pointwise Mutual Information (PMI) statistic (Church and Hanks, 1990) to compute the number of times two words co-occur within the same context (i.e. within a certain N number of words of each other in a story) relative to their overall frequency. This co-occurence measure is order-sensitive such that the first word in the pair is designated as the cause word and the second as the effect word, based on the assumption that cause events are more often described before their effects in stories, relative to the reverse. To calculate an overall causality score for two sequences $S1$ and $S2$, each cause word $c$ in $S1$ is paired with each effect word $e$ in $S2$, and the PMI scores of all word pairs are averaged: $\frac{\sum_{c \in S1} \sum_{e \in S2} PMI(c,e)}{|S1|*|S2|}$. For a given COPA item, the predicted alternative is the one that has the higher causality score with regard to the premise. Since the scores are asymmetric in assuming $S1$ is the cause of $S2$, COPA items that elicit the more plausible effect (i.e. items where "What happened as a result?" is the prompt) assign the premise and alternative to $S1$ and $S2$ respectively, whereas this assignment is reversed for items where "What was the cause of this?" is the prompt. Gordon et al. (2011) applied this approach to PMI scores taken from a corpus of one million stories extracted from personal weblogs, which were largely non-fiction stories about daily life events written from the first-person perspective. A co-occurrence between two words was counted when they appeared within 25 words of one another in the same story. This resulted in 65.2% accuracy on the COPA test set.

The PMI approach assumes a causal relation between events can be captured to some degree by their temporal co-occurrence in a story. Luo et al. (2016) introduced a variation that alternatively focuses on explicit mentions of causality in text, referred to as the CausalNet approach. They extracted sequences matching lexical templates that signify causality, e.g. $S1$ LEADS TO $S2$, $S2$ RESULTS FROM $S1$, $S2$ DUE TO $S1$, where again $S1$ is the cause event and $S2$ is the effect. As before, a co-occurrence is counted between each pair of

words ($c$, $e$) in $S1$ and $S2$, respectively. They propose a measure of *causal strength* that adapts the PMI statistic to model both necessary causality and sufficient causality for a given pair $(c, e)$. In the measure for necessary causality, a discounting factor is applied to the overall frequency of $c$ in the PMI statistic, which models the degree to which $c$ must appear in order for $e$ to appear. Alternatively, the measure for sufficient causality applies the discounting factor to the overall frequency of $e$, modeling the degree to which $c$ alone will result in the occurrence of $e$. The necessary and sufficient PMI scores for a given word pair are combined into a single causal strength score. Akin to the previous approach, the overall causality score for two sequences is given by averaging the scores for their word pairs. See Luo et al. for further technical details about the causal strength measure.

Luo et al. applied this approach to extract causal pairs from a corpus of approximately 1.6 billion web pages. They achieved 70.2% accuracy on the COPA test set, significantly outperforming the result from Gordon et al. (2011). Sasaki et al. (2017) evaluated the same CausalNet approach on a smaller corpus of web documents, ClueWeb[1], which contains 700 million pages. They discovered that treating multi-word phrases as discrete words in the pairs boosted accuracy to 71.2%. Both results indicate that causal knowledge can be extracted from large web data as an alternative to story corpora. Rather than assuming that causality is implicitly conveyed by temporally related sequences, they relied on explicit mentions of causality to filter data relevant to COPA. Still, a lot of causal knowledge in stories is not highlighted by specific lexical items. Consider the sequence "John starts a pot of coffee because he is sleepy", for example. This sequence would be extracted by the CausalNet approach since it contains one of the designated lexical markers of causality ("because"). However, the sequence "John is sleepy. He starts a pot of coffee" expresses the same causal relation but would not be captured, and we know by people's ability to answer COPA questions that they can infer this relation. Using a large web corpus can possibly compensate for missing these instances, since the same causal relations may be conveyed by sequences that contain explicit mentions of causality. However, it still means that a lot of causal information is potentially being overlooked.

[1] lemurproject.org/clueweb12/

## 4 Neural Network Approach

As mentioned in the introduction, our work initiates the exploration of neural approaches to COPA. We focus here on an encoder-decoder architecture. Originally applied to machine translation (Cho et al., 2014), encoder-decoder models have been extended to other sequence modeling tasks like dialogue generation (Serban et al., 2016; Shang et al., 2015) and poetry generation (Ghazvininejad et al., 2016; Wang et al., 2016). We propose that this technique could be similarly useful for our task in establishing a mapping between cause-effect sequence pairs. This direct modeling of co-occurrence between sequences is unique from the previous work, which relies on co-occurrence between pairs of individual words.

### 4.1 Sequence Segmentation

The inputs and outputs for the encoder-decoder model are each word sequences. Given a corpus of stories as the training set for a model, we first segmented each story by clausal boundaries. This was done heuristically by analyzing the dependency parse of each sentence. Words whose dependency label was an adverbial clause modifier (ADVCL; e.g. "After I got home, I *got* a text from her."), conjunct (CONJ; "I dropped the glass and the glass *broke*."), or prepositional complement (PCOMP; "He took me to the hospital to *seek* treatment.") were detected as the heads of clauses distinct from the main clause. All contiguous words dependent on the same head word were segmented as a separate clause. These particular labels do not capture all clausal boundaries (for example, relative clauses are not detected), but they are intended to distinguish sequences that may refer to separate narrative events (e.g. "I dropped the glass" is segmented from "and the glass broke"). This is somewhat analogous to the segmentation performed by Luo et al. (2016) that splits cause and effect clauses according to lexical templates. The difference is that the parsing labels we use for segmentation do not explicitly indicate boundaries between causally related events. We did not perform an intrinsic evaluation of this procedure in terms of how often it correctly segmented narrative events. Instead, we evaluated its impact on COPA prediction by comparing it to traditional segmentation based on sentence boundaries for the

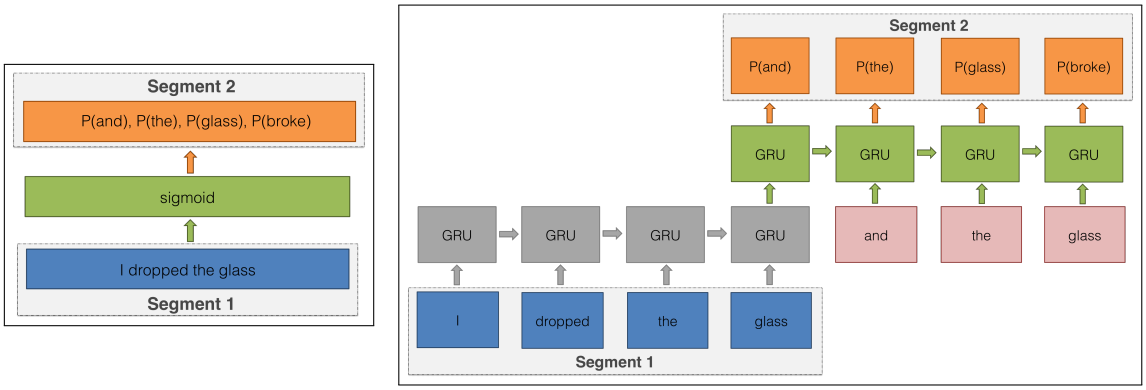


Figure 1: FFN (left) and RNN (right) encoder-decoder models

same model, as conveyed in Section 5.3.

### 4.2 Sequence Pairs

After segmenting the stories, we joined neighboring segments (i.e. clauses or sentences) into input-output segment pairs ($S1$, $S2$). In all of our experiments, we filtered pairs where one or both of the segments contained more than 20 words. We manipulated the temporal window within which these pairs were joined, by pairing all segments within N segments of each other. For a given segment at position $t$ in a story, pairs were established between all segments in $segment_t, \ldots, segment_{t+N}$. For example, when N=1, a pair was formed with the next segment only $(segment_t, segment_{t+1})$; when N=2, pairs were formed between $(segment_t, segment_{t+1})$ and $(segment_t, segment_{t+2})$. By doing this, we intended to examine the proximity of causal information in a story according to its impact on COPA prediction; we expected that more adjacent clauses would contain stronger causal relations than more distant clauses. Gordon et al. (2011) analogously evaluated this by varying the number of words within which PMI pairs were formed, but without regard to sentence or clause boundaries.

### 4.3 Encoder-decoder Models

We examined two types of encoder-decoder models: one with feed-forward (FFN) layers and one with recurrent (RNN) layers (i.e. a sequence-to-sequence model), both shown in Figure 1. In both cases, the model implicitly assumes that the input segment $S1$ represents the cause of the output segment $S2$, so the model learns to predict that $S2$ will appear as the effect of $S1$. The theoretical motivation for comparing the FFN and RNN is to determine the importance of word order for this task. The existing COPA approaches only accounted for word order to the extent of capturing word pairs within the same context of N words (though Sasaki et al. (2017) also accounted for multi-word expressions). The FFN encoder-decoder ignores word order. The model is very simple: both the input and output segments are collapsed into flat n-dimensional vectors of word counts (i.e. bag-of-words), so the hidden (encoder) layer observes all words in each segment in parallel. On the output (decoder) layer (which has sigmoid activation like the encoder), the FFN computes a score for each word indicating its probability of appearing anywhere in output segment.

In contrast, the RNN captures word order in the segments. In particular, it uses a recurrent (encoder) layer with Gated Recurrent Units (GRUs) (Cho et al., 2014) to iteratively encode the input sequence, and another recurrent (decoder) layer to represent output segment. The final hidden state of the encoder layer after observing the whole input is provided as the initial hidden state to the decoder. The decoder then iteratively computes a representation of the output sequence that is conditioned upon the input segment. For each timepoint in this decoder layer, a softmax layer is applied to predict a probability distribution over each word being observed in the segment at that particular timepoint. Both the FFN and RNN encoder-decoders are trained using the cross-entropy loss function to maximize the output word probabilities observed during training.

Once trained, a model predicts the likelihood that a cause sequence $S1$ given as input results in an effect sequence $S2$ based on the mean probability of the words in $S2$ computed by the model. When applied to COPA, consistent with the methodology described in Section 3, in items

where the prompt elicits the alternative that conveys the effect of the premise, the premise is designated as $S1$ and the alternative as $S2$. In contrast, when the item elicits the alternative describing the most likely cause of the premise, an alternative is assigned to $S1$ and the premise to $S2$. In considering the two alternatives in a COPA item, the one contained in the ($S1$, $S2$) pair that obtains the highest score is predicted as more plausible.

## 5 Initial Experiments

### 5.1 ROCStories Corpus

The PMI and CausalNet approaches to COPA made use of large web corpora. Gordon et al. (2011) proposed that stories are a rich source for the commonsense knowledge needed to answer COPA questions. Mostafazadeh et al. (2016) followed this proposal by releasing the ROCStories corpus[2], intended to be applied to commonsense reasoning tasks. The ROCStories corpus has yet to be utilized for COPA prediction. This dataset consists of 97,027 five-sentence narratives authored via crowdsourcing. In contrast to weblog stories, these stories were written with the specific objective to minimize discourse complexity and explicate prototypical causal and temporal relations between events in salient everyday scenarios. COPA items also target these latent commonsense relations, so the ROCStories appear to be particularly suitable for this domain. Table 1 shows some examples of stories in this corpus and corresponding COPA items that address the same causal knowledge. The ROCStories corpus is dramatically smaller than the datasets used in the work described in Section 3.

### 5.2 Procedure

We applied the methodology outlined in Section 4 to pairs of sequences from the ROCStories corpus. Our first set of experiments varied segmentation (clause versus sentence boundaries), distance between segments (N=1 to N=4), and the type of encoder-decoder (FFN or RNN). Note that N=4 is the maximum setting when using sentence boundaries since there are five sentences in each story, so here pairs will be formed between all sentences. For all experiments, we filtered grammatical words (i.e. all words except for adjectives, adverbs, nouns, and verbs) and lemmatized all segments, consistent with Luo et al. (2016). COPA items intentionally do not contain proper nouns, so we excluded them as well. We assembled a model lexicon that included each word occurring at least five times in the data, which totaled 9,299 words in the ROCStories. All other words were mapped to a generic <UNKNOWN> token.

The hidden layers of the FFN and RNN models each consisted of 500 dimensions. The RNN had an additional word embedding layer of 300 nodes in order to transform discrete word indices in the input segments into distributed vectors. They were both trained for 50 epochs using the Adam optimizer (Kingma and Ba, 2015) with a batch size of 100 pairs. After each epoch, we evaluated the model on the COPA development set and saved the weights that obtained the highest accuracy.

### 5.3 Results

Table 2 shows the results of these different configurations in terms of COPA accuracy. We include the results on the development set as a reference because they tended to vary from the test results. Most notably, the FFN outperformed the RNN universally, suggesting that the order of words in the segments did not provide a strong signal for prediction beyond the presence of the words themselves. Among the FFN results, the model trained on clauses with N=4 obtained the highest accuracy on the development set (66.0%), and was tied for the highest test accuracy with the model trained on clauses with N=3 (66.2%). The model with N=4 was trained on three times as many pairs as the model with N=1. We can conclude that some of these additional pairs pertained to causality, despite not appearing adjacently in the story. The impact of clause versus sentence segmentation is less clear from these results, given that the best result of 66.2% accuracy using clauses is only trivially better than the corresponding result for sentences (66.0% for N=4).

### 5.4 Other Findings

#### 5.4.1 Alternative Input Representations

In the FFN model evaluated above, the input segments were simply represented as bag-of-words vectors indicating the count of each word in the segment. Alternatively, we explored the use of pretrained word embeddings to represent the segments. We proposed that because they provide the model with some initial signal about lexical relations, the embeddings could facilitate more

[2] cs.rochester.edu/nlp/rocstories/

| ROCStories Instance | COPA Item |
|---|---|
| Susie went away to Nantucket. She wanted to relax. When she got there it was amazing. The waves were so relaxing. Susie never wanted to leave. | **Premise:** The man went away for the weekend. What was the cause of this?<br>**Alt 1*:** He wanted to relax.<br>**Alt 2:** He felt content. |
| Albert wanted to enter the spelling bee, but he was a bad speller. He practiced every day for the upcoming contest. When Albert felt that he was ready, he entered the spelling bee. In the very last round, Albert failed when he misspelled a word. Albert was very proud of himself for winning the second place trophy. | **Premise:** The girl received a trophy. What was the cause of this?<br>**Alt 1*:** She won a spelling bee.<br>**Alt 2:** She made a new friend. |
| Anna was lonely. One day, Anna went to the grocery store. Outside the store, she met a woman who was giving away kittens. Anna decided to adopt one of those kittens. Anna no longer felt lonely with her new pet. | **Premise:** The woman felt lonely. What happened as a result?<br>**Alt 1:** She renovated her kitchen.<br>**Alt 2*:** She adopted a cat. |
| April is fascinated by health and medicine. She decided to become a doctor. She studied very hard in college and medical school. April graduated at the top of her medical school class. April now works in a hospital as a doctor. | **Premise:** The woman wanted to be a doctor. What happened as a result?<br>**Alt 1:** She visited the hospital.<br>**Alt 2*:** She went to medical school. |

Table 1: Examples of stories in ROCStories corpus and similar COPA items

| | | | FFN | | RNN | |
|---|---|---|---|---|---|---|
| **Segment** | **N** | **# Pairs** | **Dev** | **Test** | **Dev** | **Test** |
| Sentence | 1 | 389,680 | 64.8 | 64.4 | 63.4 | 54.4 |
| | 2 | 682,334 | 65.2 | 65.4 | 61.2 | 57.6 |
| | 3 | 877,963 | 63.8 | 63.8 | 60.2 | 55.4 |
| | 4 | 976,568 | 63.8 | 66.0 | 59.4 | 55.6 |
| Clause | 1 | 539,342 | 64.2 | 63.6 | 59.4 | 56.8 |
| | 2 | 981,677 | 65.2 | 65.0 | 59.2 | 54.6 |
| | 3 | 1,327,010 | 65.4 | **66.2** | 63.4 | 58.0 |
| | 4 | 1,575,340 | 66.0 | **66.2** | 61.2 | 56.6 |

Table 2: Accuracy by segmentation unit and pair distance (N) for the FFN and RNN encoder-decoders trained on ROCStories

| Model | Dev | Test |
|---|---|---|
| FFN (above) | 66.0 | **66.2** |
| FFN GloVe | 65.0 | 61.6 |
| FFN ConceptNet | 61.6 | 62.4 |
| FFN Skip-thought | 66.8 | 63.8 |

Table 3: Accuracy of FFN trained on ROCStories with different input representations

specifically learning causal relations. We experimented with three sets of embedding representations. First, we encoded the words in each input segment as the sum of their GloVe embeddings[3] (Pennington et al., 2014), which represent words according to a global log-bilinear regression model trained on word co-occurrence counts in the Common Crawl corpus. We also did this using ConceptNet embeddings[4] (Li et al., 2013), which apply the word2vec skip-gram model (Mikolov et al., 2013) to tuples that specifically define commonsense knowledge relations (e.g. soak in hot-spring CAUSES get pruny skin). Lastly, we used skip-thought vectors[5] (Kiros et al., 2015), which compute one embedding representation for an entire sentence, and thus represent the sentence beyond just the sum of its individual words. Analogous to how word embedding models are trained to predict words near a given target word in a text, the skip-thought vectors represent sentences according to their relation to adjacent sentences, such that sentences with similar meanings are expected to have similar vectors. The provided skip-thought vectors are trained on the BookCorpus dataset, which is described in Section 6.

We trained the FFN model on the ROCStories with each of these three sets of embeddings. Because they obtained the best performance in the previous experiments, we configured the models to use clause segmentation and distance N=4 in constructing the pairs. Table 3 shows the results of these models, compared alongside the best result from above with the standard bag-of-words representation. Neither the GloVe nor ConceptNet embeddings performed better than the bag-of-words vectors (61.6% and 62.4% test accuracy, respectively). The skip-thought vectors performed better than bag-of-words representation on the development set (66.8%), but this improvement did not scale to the test set (63.8%).

[3] nlp.stanford.edu/projects/glove/

[4] ttic.uchicago.edu/ kgimpel/commonsense.html

[5] github.com/ryankiros/skip-thoughts

#### 5.4.2 Phrases

| Model | Dev | Test |
|---|---|---|
| FFN (above) | 66.0 | **66.2** |
| FFN Phrases | 62.6 | 64.8 |

Table 4: Accuracy of FFN trained on ROCStories with explicit phrase representations

As mentioned above, Sasaki et al. (2017) found that modeling multi-word phrases as individual words was helpful for the CausalNet approach. The RNN encoder-decoder has the opportunity to recognize phrases by modeling sequential dependencies between words, but Table 2 indicated this model was not successful relative to the FFN model. To assess whether the FFN model would benefit from phrase information, we merged all phrases in the training corpus into individual word tokens in the same manner as Sasaki et al., using their same list of phrases. We again filtered all tokens that occurred fewer than five times in the data, which resulted in the vocabulary increasing from 9,299 words to 10,694 when the phrases were included. We trained the same FFN model in Table 2 that achieved the best result (clause segmentation, N=4, and bag-of-words input representation). The test accuracy, relayed for clarity in Table 4 alongside the above best result, was 64.8%, indicating there was no benefit to modeling phrases in this particular configuration.

#### 5.4.3 Comparison with Existing Approaches

| Model | Dev | Test |
|---|---|---|
| FFN (above) | 66.0 | **66.2** |
| PMI | 60.0 | 62.4 |
| CausalNet | 50.2 | 51.8 |

Table 5: Accuracy of PMI and CausalNet trained on ROCStories

To establish a comparison between our encoder-decoder approach and the existing models applied to the same dataset, we trained the PMI model on the ROCStories. Rather than using a fixed word window, we computed the PMI scores for all words in each story, which generally corresponds to using distance N=4 among sentence segments in the encoder-decoder. Table 5 shows that this approach had 62.4% test accuracy, so our new approach outperformed it on this particular dataset.

For completeness, we also applied the CausalNet approach to this dataset. Its poor performance (51.8%) is unsurprising, because the lexical templates used to extract causal pairs only matched 4,964 sequences in the ROCStories. This demonstrates that most of the causal information contained in these stories is conveyed implicitly.

## 6 Experiments on Other Datasets

Gordon et al. (2011) found that the PMI approach trained on blog stories performed better on COPA than the same model trained on books in Project Gutenburg[6], despite the much larger size of the latter. Beyond this, there has been limited exploration of the impact of different training datasets on COPA prediction, so we were motivated to examine this. Thus, we applied the FFN encoder-decoder approach to the following datasets:

**Visual Storytelling (VIST):** 50,200 five-sentence stories[7] authored through crowdsourcing in support of research on vision-to-language tasks (Huang et al., 2016). Participants were prompted to write a story from a sequence of photographs depicting salient "storyable" events.

**CNN/DailyMail corpus:** 312,085 bullet-item summaries[8] of news articles, which have been used for work on reading comprehension and summarization (Chen et al., 2016; See et al., 2017).

**CMU Book/Movie Plot Summaries (CMU Plots):** 58,862 plot summaries[9] from Wikipedia, which have been used for story modeling tasks like inferring relations between story characters (Bamman et al., 2014; Srivastava et al., 2016).

**BookCorpus:** 8,032 self-published fiction novels, a subset of the full corpus[10] of 11,000 books.

**Blog Stories:** 1 million weblog stories used in the COPA experiments by Gordon et al. (2011) identified above.

**ClueWeb Pairs:** Approximately 150 million sequence pairs extracted from the ClueWeb corpus by Sasaki et al. (2017) using the CausalNet lexical templates method.

### 6.1 Procedure and Results

We trained the FFN model with the best-performing configuration from the ROCStories experiments (clause segments, N=4, bag-of-words input). After determining that the lexicon used in the previous experiments included most of the words (93.5%) in the COPA development set, we re-used this same lexicon to avoid the inefficiency of assembling a new one for each separate corpus. We also trained a model on the initial 45,502 stories in the ROCStories (ROCStories-Half) to further analyze the impact of this dataset.

| Dataset | # Pairs | Dev | Test |
|---|---|---|---|
| ROCStories-Half | 762,130 | 64.0 | 62.6 |
| VIST | 854,810 | 58.2 | 49.2 |
| ROCStories-Full | 1,575,340 | 66.0 | **66.2** |
| CNN/DailyMail | 3,255,010 | 59.4 | 51.8 |
| CMU Plots | 6,094,619 | 57.8 | 51.0 |
| ClueWeb Pairs | 157,426,812 | 60.8 | 61.2 |
| Blog Stories | 222,564,571 | 58.4 | 57.2 |
| BookCorpus | 310,001,015 | 58.2 | 55.0 |

Table 6: Accuracy of the FFN encoder-decoder on different datasets

Table 6 shows the results for these datasets compared alongside the ROCStories result from above (ROCStories-Full), listed in ascending order of the number of training pairs they contain. As shown, none of the other datasets reach the level of accuracy of ROCStories-Full (66.2%). Even the model trained on only the initial half of this corpus outperforms the others (62.6%). The next closest result is for the ClueWeb Pairs, which had 61.2% test accuracy despite containing 100 times more pairs than the ROCStories. The larger Blog Stories and BookCorpus datasets did not have much impact, despite that the Blog Stories obtained 65.2% accuracy in the PMI approach. One speculative explanation for this is that our approach is highly dependent on the *density* of COPA-relevant knowledge contained in a dataset. As mentioned above, authors of the ROCStories were instructed to emphasize the most obvious possibilities for 'what happens next' in prototypical scenarios. These expectations align with the correct COPA alternatives. However, naturally occurring stories often focus on events that violate commonsense expectations, since these events make for more salient stories (Schank and Ableson, 1995). Thus, they may show greater diversity in 'what happens next' relative to the ROCStories. This diversity was seemingly more distracting for our encoder-decoder architecture than for the ex-

[6] gutenberg.org/
[7] visionandlanguage.net/VIST/
[8] github.com/danqi/rc-cnn-dailymail
[9] cs.cmu.edu/~ark/personas/; cs.cmu.edu/~dbamman/booksummaries.html
[10] yknzhu.wixsite.com/mbweb

isting approaches. Accordingly, despite all being related to narrative, the VIST, CNN/DailyMail, and CMU Plots datasets were also ineffective on the test set with regard to this model.

## 7 Conclusion

In summary, we pursued a neural encoder-decoder approach for predicting causally related events in the COPA framework. To our knowledge this is the first work to evaluate a neural-based model for this task. Our best result obtained 66.2% accuracy. This is lower than the current state-of-the-art of 71.2%, but our experiments motivate some opportunities for future work. We demonstrated the usefulness of the ROCStories for this task, as our model appeared to benefit from its density of commonsense knowledge. The gap between 66.2% and 71.2% is not dramatic in light of the massive size advantage of the data used to obtain the latter result. However, the ROCStories corpus is a crowdsourced dataset and thus will not grow naturally over time like web data, so it may not be practical to rely exclusively on this type of specially authored resource either. The CausalNet approach proposed a useful way to isolate commonsense knowledge in generic text by relying on causal cues, but because many causal relations are not marked by specific lexical items, it still overlooks a lot that is relevant to COPA. On the other hand, not all temporally related events in a story are causally related. Because we did not make this distinction, some of the pairs we modeled were likely not indicative of causality and thus may not have contributed accurately to COPA prediction. Research on automatically detecting more latent linguistic features specifically associated with the expression of causal knowledge in text would likely have a large impact on this endeavor.

## Acknowledgments

We would like to thank the authors of Sasaki et al. (2017) for sharing the data and resources associated with their work.

The projects or efforts depicted were or are sponsored by the U.S. Army. The content or information presented does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

## References

David Bamman, Brendan O'Connor, and Noah A Smith. 2014. Learning latent personas of film characters. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. page 352.

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

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Association for Computational Linguistics (ACL)*.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation* page 103.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16(1):22–29.

Natalie Dehn. 1981. Story Generation After TALE-SPIN. *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81)* pages 16–18.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *EMNLP*. pages 1183–1191.

Andrew Gordon, Cosmin Adrian Bejan, and Kenji Sagae. 2011. Commonsense Causal Reasoning Using Millions of Personal Stories. *Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)* pages 1180–1185.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations (ICLR)*. San Diego.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., pages 3294–3302.

Michael Lebowitz. 1985. Story-telling as planning and learning. *Poetics* 14(6):483–502.

Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. Story Generation with Crowdsourced Plot Graphs. In *27th AAAI Conference on Artificial Intelligence*.

Zhiyi Luo, Yuchen Sha, Kenny Q. Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *15th International Conference on Principles of Knowledge Representation and Reasoning (KR-2016)*.

James R Meehan. 1977. TALE-SPIN, An Interactive Program that Writes Stories. In *5th International Joint Conference on Artificial Intelligence*. pages 91–98.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*. pages 839–849.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of Plausible Alternatives : An Evaluation of Commonsense Causal Reasoning. *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning* pages 90–95.

Shota Sasaki, Sho Takase, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2017. Handling multiword expressions in causality estimation. In *IWCS 201712th International Conference on Computational SemanticsShort papers*.

Roger C Schank and Robert P Abelson. 1977. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Roger C Schank and Robert P Ableson. 1995. Knowledge and memory: The real story. *In. Rober S. Wyer (Ed.), Knowledge and memory: The real story* pages 1–85.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Association for Computational Linguistics (ACL)*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*. pages 3776–3784.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Proceedings of the 53th Annual Meeting of Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP'15)*. pages 1577–1586.

Shashank Srivastava, Snigdha Chaturvedi, and Tom M Mitchell. 2016. Inferring interpersonal relations in narrative summaries. In *AAAI*. pages 2807–2813.

Qixin Wang, Tianyi Luo, Dong Wang, and Chao Xing. 2016. Chinese song iambics generation with neural attention-based model. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*. pages 2943–2949.