Quality Signals in Generated Stories

Manasvi Sagarkar[*] **John Wieting**[†] **Lifu Tu**[‡] **Kevin Gimpel**[‡]

[*]University of Chicago, Chicago, IL, 60637, USA
[†]Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[‡]Toyota Technological Institute at Chicago, Chicago, IL, 60637, USA

manasvi@uchicago.edu, jwieting@cs.cmu.edu, {lifu,kgimpel}@ttic.edu

Abstract

We study the problem of measuring the quality of automatically-generated stories. We focus on the setting in which a few sentences of a story are provided and the task is to generate the next sentence ("continuation") in the story. We seek to identify what makes a story continuation interesting, relevant, and have high overall quality. We crowdsource annotations along these three criteria for the outputs of story continuation systems, design features, and train models to predict the annotations. Our trained scorer can be used as a rich feature function for story generation, a reward function for systems that use reinforcement learning to learn to generate stories, and as a partial evaluation metric for story generation.

1 Introduction

We study the problem of automatic story generation in the climate of neural network natural language generation methods. Story generation (Mani, 2012; Gervás, 2012) has a long history, beginning with rule-based systems in the 1970s (Klein et al., 1973; Meehan, 1977). Most story generation research has focused on modeling the plot, characters, and primary action of the story, using simplistic methods for producing the actual linguistic form of the stories (Turner, 1993; Riedl and Young, 2010). More recent work learns from data how to generate stories holistically without a clear separation between content selection and surface realization (McIntyre and Lapata, 2009), with a few recent methods based on recurrent neural networks (Roemmele and Gordon, 2015; Huang et al., 2016).

We follow the latter style and focus on a setting in which a few sentences of a story are provided (the **context**) and the task is to generate the next sentence in the story (the **continuation**). Our goal is to produce continuations that are both interesting and relevant given the context.

Neural networks are increasingly employed for natural language generation, most often with encoder-decoder architectures based on recurrent neural networks (Cho et al., 2014; Sutskever et al., 2014). However, while neural methods are effective for generation of individual sentences conditioned on some context, they struggle with coherence when used to generate longer texts (Kiddon et al., 2016). In addition, it is challenging to apply neural models in less constrained generation tasks with many valid solutions, such as open-domain dialogue and story continuation.

The story continuation task is difficult to formulate and evaluate because there can be a wide variety of reasonable continuations for typical story contexts. This is also the case in open-domain dialogue systems, in which common evaluation metrics like BLEU (Papineni et al., 2002) are only weakly correlated with human judgments (Liu et al., 2016). Another problem with metrics like BLEU is the dependence on a gold standard. In story generation and open-domain dialogue, there can be several equally good continuations for any given context which suggests that the quality of a continuation should be computable without reliance on a gold standard.

In this paper, we study the question of identifying the characteristics of a good continuation for a given context. We begin by building several story generation systems that generate a continuation from a context. We develop simple systems based on recurrent neural networks and similarity-based retrieval and train them on the ROC story dataset (Mostafazadeh et al., 2016). We use crowdsourcing to collect annotations of the quality of the continuations without revealing the gold standard. We ask annotators to judge continuations along three distinct criteria: overall quality, relevance, and interestingness. We collect multiple annotations for 4586 context/continuation pairs. These annotations permit us to compare methods for story generation and to study the relationships among the criteria. We analyze our annotated dataset by developing features of the context and continuation and measuring their correlation with each criterion.

We combine these features with neural networks to build models that predict the human scores, thus attempting to automate the process of human quality judgment. We find that our predicted scores correlate well with human judgments, especially when using our full feature set. Our scorer can be used as a rich feature function for story generation or a reward function for systems that use reinforcement learning to learn to generate stories. It can also be used as a partial evaluation metric for story generation.[1] Examples of contexts, generated continuations, and quality predictions from our scorer are shown in Table 3. The annotated data and trained scorer are available at the authors' websites.

2 Related Work

Research in automatic story generation has a long history, with early efforts driven primarily by hand-written rules (Klein et al., 1973; Meehan, 1977; Dehn, 1981; Lebowitz, 1985; Turner, 1993), often drawing from theoretical analysis of stories (Propp, 1968; Schank and Abelson, 1975; Thorndyke, 1977; Wilensky, 1983). Later methods were based on various methods of planning from artificial intelligence (Theune et al., 2003; Oinonen et al., 2006; Riedl and Young, 2010) or commonsense knowledge resources (Liu and Singh, 2002; Winston, 2014). A detailed summary of this earlier work is beyond our scope; for surveys, please see Mani (2012), Gervás (2012), or Gatt and Krahmer (2017).

More recent work in story generation has focused on data-driven methods (McIntyre and Lapata, 2009, 2010; McIntyre, 2011; Elson, 2012; Daza et al., 2016; Roemmele, 2016). The generation problem is often constrained via anchoring to some other input, such as a topic or list of keywords (McIntyre and Lapata, 2009), a sequence of images (Huang et al., 2016), a set of loosely-connected sentences (Jain et al., 2017), or settings in which a user and agent take turns adding sentences to a story (Swanson and Gordon, 2012; Roemmele and Gordon, 2015; Roemmele, 2016).

Our annotation criteria—relevance, interestingness, and overall quality—are inspired by those from prior work. McIntyre and Lapata (2009) similarly obtain annotations for story interestingness. They capture coherence in generated stories by using an automatic method based on sentence shuffling. We discuss the relationship between relevance and coherence below in Section 3.2.

Roemmele et al. (2017) use automated linguistic analysis to evaluate story generation systems. They explore the various factors that affect the quality of a story by measuring feature values for different story generation systems, but they do not obtain any quality annotations as we do here.

Since there is little work in automatic evaluation of story generation, we can turn to the related task of open-domain dialogue. Evaluation of dialogue systems often uses perplexity or metrics like BLEU (Papineni et al., 2002), but Liu et al. (2016) show that most common evaluation metrics for dialog systems are correlated very weakly with human judgments. Lowe et al. (2017) develop an automatic metric for dialog evaluation by training a model to predict crowdsourced quality judgments. While this idea is very similar to our work, one key difference is that their annotators were shown both system outputs and the gold standard for each context. We fear this can bias the annotations by turning them into a measure of similarity to the gold standard, so we do not show the gold standard to annotators.

Wang et al. (2017) use crowdsourcing (upvotes on Quora) to obtain quality judgments for short stories and train models to predict them. One difference is that we obtain annotations for three distinct criteria, while they only use upvotes. Another difference is that we collect annotations for both manually-written continuations and a range of system-generated continuations, with the goal of using our annotations to train a scorer that can be used within training.

3 Data Collection

Our goal is to collect annotations of the quality of a sentence in a story given its preceding sentences. We use the term **context** to refer to the preceding sentences and **continuation** to refer to the next

[1] However, since our scorer does not use a gold standard, it is possible to "game" the metric by directly optimizing the predicted score, so if used as an evaluation metric, it should still be validated with a small-scale manual evaluation.

sentence being generated and evaluated. We now describe how we obtain ⟨context, continuation⟩ pairs from automatic and human-written stories for crowdsourcing quality judgments.

We use the ROC story corpus (Mostafazadeh et al., 2016), which contains 5-sentence stories about everyday events. We use the initial data release of 45,502 stories. The first 45,002 stories form our training set (TRAIN) for story generation models and the last 500 stories form our development set (DEV) for tuning hyperparameters while training story generation models. For collecting annotations, we compile a dataset of 4586 context-continuation pairs, drawing contexts from DEV as well as the 1871-story validation set from the ROC Story Cloze task (Mostafazadeh et al., 2016).

For contexts, we use 3- and 4-sentence prefixes from the stories in this set of 4586. We use both 3 and 4 sentence contexts as we do not want our annotated dataset to include only story endings (for the 4-sentence contexts, the original 5th sentence is the ending of the story) but also more general instances of story continuation. We did not use 1 or 2 sentence contexts because we consider the space of possible continuations for these short contexts to be too unconstrained and thus it would be difficult for both systems and annotators.

We generated continuations for each context using a variety of systems (described in Section 3.1) as well as simply taking the human-written continuation from the original story. We then obtained annotations for the continuation with its context via crowdsourcing, described in Section 3.2.

3.1 Story Continuation Systems

In order to generate a dataset with a range of qualities, we consider six ways of generating the continuation of the story, four based on neural sequence-to-sequence models and two using human-written sentences. To lessen the possibility of annotators seeing the same context multiple times, which could bias the annotations, we used at most two methods out of six for generating the continuation for a particular context.

3.1.1 Sequence-to-Sequence Models

We used a standard sequence-to-sequence (SEQ2SEQ) neural network model (Sutskever et al., 2014) to generate continuations given contexts. We trained the models on TRAIN and tuned on DEV. We generated 180,008 ⟨context, continuation⟩ pairs from TRAIN, where the continuation is always a single sentence and the context consists of all previous sentences in the story. We trained a 3-layer bidirectional SEQ2SEQ model, with each layer having hidden vector dimensionality 1024. The size of the vocabulary was 31,220. We used scheduled sampling (Bengio et al., 2015), using the previous ground truth word in the decoder with probability 0.5^t, where t is the index of the mini-batch processed during training. We trained the model for 20,000 epochs with a batch size of 100. We began training the model on consecutive sentence pairs (so the context was only a single sentence), then shifted to training on full story contexts.

We considered four different methods for the decoding function of our SEQ2SEQ model:

- SEQ2SEQ-GREEDY: return the highest-scoring output under greedy ($\arg\max$) decoding.
- SEQ2SEQ-DIV: return the kth-best output using a diverse beam search (Vijayakumar et al., 2016) with beam size $k = 10$.
- SEQ2SEQ-SAMPLE: sample words from the distribution over output words at each step using a temperature parameter $\tau = 0.4$.
- SEQ2SEQ-REVERSE: reverse input sequence (at test time only) and use greedy decoding.

Each decoding rule contributes one eighth of the total data generated for annotation, so the SEQ2SEQ models account for one half of the ⟨context, continuation⟩ pairs to be annotated.

3.1.2 Human Generated Outputs

For human generated continuations, we use two methods. The first is simply the gold standard continuation from the ROC stories dataset, which we call HUMAN. The second finds the most similar context in the ROC training corpus, then returns the continuation for that context. To compute similarity between contexts, we use the sum of two similarity scores: BLEU score (Papineni et al., 2002) and the overall sentence similarity described by Li et al. (2006). Since this method is similar to an information retrieval-based story generation system, we refer to it as RETRIEVAL. HUMAN and RETRIEVAL each contribute a fourth of the total data generated for annotation.

3.2 Crowdsourcing Annotations

We used Amazon Mechanical Turk to collect annotations of continuations paired with their contexts. We collected annotations for 4586 context-continuation pairs, collecting the following three criteria for each pair:

- **Overall quality (O)**: a subjective judgment by the annotator of the quality of the continuation, i.e., roughly how much the annotator thinks the continuation adds to the story.
- **Relevance (R)**: a measure of how relevant the continuation is to the context. This addresses the question of whether the continuation fits within the world of the story.
- **Interestingness (I)**: a measure of the amount of new (but still relevant) information added to the story. We use this to measure whether the continuation makes the story more interesting.

Our criteria follow McIntyre and Lapata (2009) who used interestingness and coherence as two quality criteria for story generation. Our notion of relevance is closely related to coherence; when thinking of judging a continuation, we believed that it would be more natural for annotators to judge the relevance of the continuation to its context, rather than judging the coherence of the resulting story. That is, coherence is a property of a discourse, while relevance is a property of a continuation (in relation to the context).

Our overall quality score was intended to capture any remaining factors that determine human quality judgment. In preliminary annotation experiments, we found that the overall score tended to capture a notion of fluency/grammaticality, hence we decided not to annotate this criterion separately. We asked annotators to forgive minor ungrammaticalities in the continuations and rate them as long as they could be understood. If annotators could not understand the continuation, we asked them to assign a score of 0 for all criteria.

We asked the workers to rate the continuations on a scale of 1 to 10, with 10 being the highest score. We obtained annotations from two distinct annotators for each pair and for each criterion, adding up to a total of $4586 \times 2 \times 3 = 27516$ judgments. We asked annotators to annotate all three criteria for a given pair simultaneously in one HIT.[2] We required workers to be located in the United States, to have a HIT approval rating greater than 97%, and to have had at least 500 HITs approved. We paid $0.08 per HIT. Since task duration can be difficult to estimate from HIT times (due to workers becoming distracted or working on multiple HITs simultaneously), we report the top 5 modes of the time duration data in seconds. For pairs with 3 sentences in the context, the most frequent durations are 11, 15, 14, 17, and 21 seconds. For 4 sentences, the most frequent durations are 18, 20, 19, 21, and 23 seconds.

Criterion	Mean	Std.	IA MAD	IA SDAD
Overall	5.2	2.5	2.1	1.6
Relevance	5.2	3.0	2.3	1.8
Interestingness	4.6	2.5	2.1	1.9

Table 1: Means and standard deviations for each criterion, as well as inter-annotator (IA) mean absolute differences (MAD) and standard deviations of absolute differences (SDAD).

We required each worker to annotate no more than 150 continuations so as not to bias the data collected. After collecting all annotations, we adjusted the scores to account for how harshly or leniently each worker scored the sentences on average. We did this by normalizing each score by the absolute value of the difference between the worker's mean score and the average mean score of all workers for each criterion. We only normalized scores of workers who annotated more than 10 pairs in order to ensure reliable worker means. We then averaged the two adjusted sets of scores for each pair to get a single set of scores.

4 Dataset Analysis

Table 1 shows means and standard deviations for the three criteria. The means are similar across the three, though interestingness has the lowest, which aligns with our expectations of the ROC stories. For measuring inter-annotator agreement, we consider the mean absolute difference (MAD) of the two judgments for each pair.[3] Table 1 shows the MADs for each criterion and the corresponding standard deviations (SDAD). Overall quality and interestingness showed slightly lower MADs than relevance, though all three criteria are similar.

The average scores for each data source are shown in Table 2. The ranking of the systems is

[2] In a preliminary study, we experimented with asking for each criterion separately to avoid accidental correlation of the criteria, but found that it greatly reduced cumulative cognitive load for each annotator to do all three together.

[3] Cohen's Kappa is not appropriate for our data because, while we obtained two annotations for each pair, they were not always from the same pair of annotators. In this case, an annotator-agnostic metric like MAD (and its associated standard deviation) is a better measure of agreement.

System	#	O	R	I
SEQ2SEQ-GREEDY	596	4.18	4.09	3.81
SEQ2SEQ-DIV	584	3.36	3.50	3.00
SEQ2SEQ-SAMPLE	578	3.69	3.70	3.42
SEQ2SEQ-REVERSE	577	4.61	4.39	4.02
RETRIEVAL	1086	5.68	4.93	5.15
HUMAN	1165	7.22	8.05	6.33

Table 2: Average criteria scores for each system (O = overall, R = relevance, I = interestingness).

consistent across criteria. Human-written continuations are best under all three criteria. The HUMAN relevance average is higher than interestingness. This matches our intuitions about the ROC corpus: the stories were written to capture commonsense knowledge about everyday events rather than to be particularly surprising or interesting stories in their own right. Nonetheless, we do find that the HUMAN continuations have higher interestingness scores than all automatic systems.

The RETRIEVAL system actually outperforms all SEQ2SEQ systems on all criteria, though the gap is smallest on relevance. We found that the SEQ2SEQ systems often produced continuations that fit topically within the world suggested by the context, though they were often generic or merely topically relevant without necessarily moving the story forward. We found S2S-GREEDY produced outputs that were grammatical and relevant but tended to be more mundane whereas S2S-REVERSE tended to produce slightly more interesting outputs that were still grammatical and relevant on average. The sampling and diverse beam search outputs were frequently ungrammatical and therefore suffer under all criteria.

We show sample outputs from the different systems in Table 3. We also show predicted criteria scores from our final automatic scoring model (see Section 6 for details). We show predicted rather than annotated scores here because for a given context, we did not obtain annotations for all continuations for that context. We can see some of the characteristics of the different models and understand how their outputs differ. The RETRIEVAL outputs are sometimes more interesting than the HUMAN outputs, though they often mention new entities that were not contained in the context, or they may be merely topically related to the context without necessarily resulting in a coherent story. This affects interestingness as well, as a continuation must first be relevant in order to be interesting.

4.1 Relationships Among Criteria

Table 4 shows correlations among the criteria for different sets of outputs. RETRIEVAL outputs show a lower correlation between overall score and interestingness than HUMAN outputs. This is likely because the RETRIEVAL outputs with high interestingness scores frequently contained more surprising content such as new character names or new actions/events that were not found in the context. Therefore, a high interestingness score was not as strongly correlated with overall quality as with HUMAN outputs, for which interesting continuations were less likely to contain erroneous new material.

HUMAN continuations have a lower correlation between relevance and interestingness than the RETRIEVAL or SEQ2SEQ models. This is likely because nearly all HUMAN outputs are relevant, so their interestingness does not depend on their relevance. For SEQ2SEQ, the continuations can only be interesting if they are first somewhat relevant to the context; nonsensical output was rarely annotated as interesting. Thus the SEQ2SEQ relevance and interestingness scores have a higher correlation than for HUMAN or RETRIEVAL.

The lower rows show correlations for different levels of overall quality. For stories whose overall quality is greater than 7.5, the correlations between the overall score and the other two criteria is higher than when the overall quality is lower. The correlation between relevance and interestingness is not as high (0.34). The stories at this quality level are already at least somewhat relevant and understandable, hence like HUMAN outputs, the interestingness score is not as dependent on the relevance score. For stories with overall quality below 2.5, the stories are often not understandable so annotators assigned low scores to all three criteria, leading to higher correlation among them.

4.2 Features

We also analyze our dataset by designing features of the ⟨context, continuation⟩ pair and measuring their correlation with each criterion.

4.2.1 Shallow Features

We consider simple features designed to capture surface-level characteristics of the continuation:

- **Length**: number of tokens in the continuation.
- **Relative length**: the length of the continuation divided by the length of the context.

<table>
<tr><td colspan="5"><b>Context 1:</b> Tripp wanted to learn how to put a topspin on his serve . He was a more advanced tennis player . He sought out a tennis pro to help him . He finally perfected his topspin .</td></tr>
<tr><th>System</th><th>Continuation</th><th>O</th><th>R</th><th>I</th></tr>
<tr><td>S2S-Greedy</td><td>He won the game .</td><td>4.12</td><td>4.99</td><td>3.45</td></tr>
<tr><td>S2S-Div</td><td>Now he had the game.</td><td>4.25</td><td>5.04</td><td>3.60</td></tr>
<tr><td>S2S-Sample</td><td>Now , he is able to play his .</td><td>4.48</td><td>4.88</td><td>3.94</td></tr>
<tr><td>S2S-Reverse</td><td>He took a few minutes .</td><td>4.92</td><td>5.77</td><td>4.22</td></tr>
<tr><td>Retrieval</td><td>Lyn now has a great backhand .</td><td>6.05</td><td>7.14</td><td>5.34</td></tr>
<tr><td>Human</td><td>His game improved even more.</td><td>5.05</td><td>6.16</td><td>4.36</td></tr>
<tr><td colspan="5"><b>Context 2:</b> Neil had just entered the country of Oman . He found the desert land to be enchanting . The women dressed beautifully and the men were friendly . Neil felt very comfortable in Oman .</td></tr>
<tr><td>S2S-Greedy</td><td>Neil decided to try the best man.</td><td>5.47</td><td>6.13</td><td>4.74</td></tr>
<tr><td>S2S-Div</td><td>They were days and decided</td><td>4.66</td><td>5.65</td><td>3.96</td></tr>
<tr><td>S2S-Sample</td><td>Neil Neil the trip trip of the trip of the trip</td><td>4.95</td><td>5.66</td><td>4.72</td></tr>
<tr><td>S2S-Reverse</td><td>He took a tour of the city.</td><td>3.97</td><td>4.83</td><td>3.64</td></tr>
<tr><td>Retrieval</td><td>Neil saw that South Koreans were a very kind people !</td><td>6.26</td><td>6.94</td><td>5.66</td></tr>
<tr><td>Human</td><td>He wished he could stay forever!</td><td>6.24</td><td>7.22</td><td>5.58</td></tr>
<tr><td colspan="5"><b>Context 3:</b> Ed and Emma were twins and wanted to have matching Halloween costumes . But they couldn ’t agree on a costume ! Ed wanted to be a superhero and Emma wanted to be a mermaid .</td></tr>
<tr><td>S2S-Greedy</td><td>He took out and could make to work .</td><td>4.60</td><td>5.11</td><td>4.11</td></tr>
<tr><td>S2S-Div</td><td>So , s ’ and they would learn .</td><td>4.71</td><td>5.41</td><td>4.18</td></tr>
<tr><td>S2S-Sample</td><td>They decided went their great time and they their family . s house .</td><td>4.86</td><td>5.50</td><td>4.58</td></tr>
<tr><td>S2S-Reverse</td><td>They decided to try to their local home .</td><td>4.74</td><td>5.21</td><td>4.22</td></tr>
<tr><td>Retrieval</td><td>Then their mom offered a solution to please them both .</td><td>5.59</td><td>6.11</td><td>5.05</td></tr>
<tr><td>Human</td><td>Then their mom said she could make costumes that ’d please them both .</td><td>6.17</td><td>6.71</td><td>5.69</td></tr>
</table>

Table 3: Sample system outputs for different contexts. Final three columns show *predicted* scores from our trained scorer (see Section 6 for details).

	Corr(O,R)	Corr(O,I)	Corr(R,I)
Human	0.70	0.63	0.44
Retrieval	0.68	0.52	0.47
Human + Ret.	0.76	0.61	0.53
Seq2Seq-All	0.72	0.70	0.59
Overall > 7.5	0.46	0.47	0.34
5 < Overall < 7.5	0.44	0.31	0.24
2.5 < Overall < 5	0.38	0.35	0.38
Overall < 2.5	0.41	0.41	0.38
Overall > 2.5	0.76	0.69	0.59

Table 4: Pearson correlations between criteria for different subsets of the annotated data.

- **Language model**: perplexity from a 4-gram language model with modified Kneser-Ney smoothing estimated using KenLM (Heafield, 2011) from the Personal Story corpus (Gordon and Swanson, 2009), which includes about 1.6 million personal stories from weblogs.
- **IDF**: the average of the inverse document frequencies (IDFs) across all tokens in the continuation. The IDFs are computed using Wikipedia sentences as “documents”.

4.2.2 PMI Features

We use features based on pointwise mutual information (PMI) of word pairs in the context and continuation. We take inspiration from methods developed for the Choice of Plausible Alternatives (COPA) task (Roemmele et al., 2011), in which a premise is provided with two alternatives. Gordon et al. (2011) obtained strong results by using PMIs to compute a score that measures the causal relatedness between a premise and its potential alternatives. For a ⟨context, continuation⟩ pair, we compute the following score (Gordon et al., 2011):

$$s_{\text{pmi}} = \frac{\sum_{u \in \text{context}} \sum_{v \in \text{continuation}} \text{PMI}(u, v)}{N_{\text{context}} N_{\text{contination}}}$$

where N_{context} and $N_{\text{continuation}}$ are the numbers of tokens in the context and continuation. We create 6 versions of the above score, combining three window sizes (10, 25, and 50) with both standard PMI and positive PMI (PPMI). To compute PMI/PPMI, we use the Personal Story corpus.[4] For efficiency and robustness, we only compute PMI/PPMI of a word pair if the pair appears more than 10 times in the corpus using the particular window size.

4.2.3 Entity Mention Features

We compute several features to capture how relevant the continuation is to the input. In

[4] We use Wikipedia for IDFs and the Personal Story corpus for PMIs. IDF is a simpler statistic which is presumed to be similar across a range of large corpora for most words; we use Wikipedia because it has broad coverage in terms of vocabulary. PMIs require computing word pair statistics and are therefore expected to be more data-dependent, so we chose the Personal Story corpus due to its effectiveness for related tasks (Gordon et al., 2011).

Feature	O	R	I
Length	0.007	0.055	0.071
Relative length	0.018	0.020	0.060
Language model	0.025	0.034	0.058
IDF	0.418	0.316	0.408
PPMI ($w = 10$)	0.265	0.321	0.224
PPMI ($w = 25$)	0.289	0.341	0.249
PPMI ($w = 50$)	0.299	0.351	0.259
Has old mentions	0.050	0.151	0.023
Number of old mentions	0.057	0.146	0.049
Has new mentions	-0.048	-0.115	-0.026
Number of new mentions	-0.052	-0.119	-0.029
Has new names	-0.005	-0.129	0.017
Number of new names	-0.005	-0.130	0.017
Is HUMAN?	*0.56*	*0.62*	*0.50*
Is HUMAN ∪ RETRIEVAL?	*0.60*	*0.49*	*0.56*

Table 5: Spearman correlations between features and annotations. The final two rows are "oracle" binary features that return 1 for continuations from those sets.

order to compute these features we use the part-of-speech tagging, named entity recognition (NER), and coreference resolution tools in Stanford CoreNLP (Manning et al., 2014):

- **Has old mentions**: a binary feature that returns 1 if the continuation has "old mentions," i.e., mentions that are part of a coreference chain that began in the context.
- **Number of old mentions**: the number of old mentions in the continuation.
- **Has new mentions**: a binary feature that returns 1 if the continuation has "new mentions," i.e., mentions that are not part of any coreference chain that began in the context.
- **Number of new mentions**: the number of new mentions in the continuation.
- **Has new names**: if the continuation has new mentions, this binary feature returns 1 if any of the new mentions is a name, i.e., if the mention is a person named entity from the NER system.
- **Number of new names**: the number of new names in the continuation.

4.3 Comparing Features

Table 5 shows Spearman correlations between our features and the criteria.[5] The length features have small positive correlations with all three criteria, showing highest correlation with interestingness. Language model perplexity shows weak correlation for all three measures, with its highest correlation for interestingness. The SEQ2SEQ models output very common words which lets them have relatively low perplexities even with occasional disfluencies, while the human-written outputs contain more rare words.

The IDF feature shows highest correlation with overall and interestingness, and lower correlation with relevance. This is intuitive since the IDF feature will be largest when many rare words are used, which is expected to correlate with interestingness more than relevance. We suspect IDF correlates so well with overall because SEQ2SEQ models typically generate common words, so this feature may partially separate the SEQ2SEQ from HUMAN/RETRIEVAL.

Unlike IDF, the PPMI scores (with window sizes w shown in parentheses) show highest correlations with relevance. This is intuitive, since PPMI will be highest when topical coherence is present in the discourse. Higher correlations are found when using larger window sizes.[6]

The old mentions features have the highest correlation with relevance, as expected. A continuation that continues coreference chains is more likely to be relevant. The new mention/name features have negative correlations with relevance, which is also intuitive: introducing new characters makes the continuation less relevant.

To explore the question of separability between machine and human-written continuations, we measured correlations of "oracle" features that simply return 1 if the output was generated by humans and 0 if it was generated by a system. Such features are highly correlated with all three criteria as seen in the final two rows of Table 5. This suggests that human annotators strongly preferred human generated stories over our models' outputs. Some features may correlate with the annotated criteria if they separate human- and machine-generated continuations (e.g., IDF).

5 Methods for Score Prediction

We now consider ways to build models to predict our criteria. We define neural networks that take as input representations of the context/continuation pair $\langle b, c\rangle$ and our features and output a continuous value for each predicted criterion.

We experiment with two ways of representing the input based on the embeddings of b and c,

[5] These use the combined training and validation sets; we describe splitting the data below in Section 6.

[6] We omit full results for brevity, but the PPMI features showed slightly higher correlations than PMI features.

which we denote $\mathbf{v}_b$ and $\mathbf{v}_c$ respectively. The first ("cont") uses only the continuation embedding without any representation of the context or the similarity between the context and continuation: $\mathbf{x}_{\text{cont}} = \langle \mathbf{v}_c \rangle$. The second ("sim+cont") also contains the elementwise multiplication of the context and continuation embeddings concatenated with the absolute difference: $\mathbf{x}_{\text{sim+cont}} = \langle \mathbf{v}_b \odot \mathbf{v}_c, |\mathbf{v}_b - \mathbf{v}_c|, \mathbf{v}_c \rangle$.

To compute representations $\mathbf{v}$, we use the average of character n-gram embeddings (Huang et al., 2013; Wieting et al., 2016), fixing the output dimensionality to 300. We found this to outperform other methods. In particular, the next best method used gated recurrent averaging networks (GRANs; Wieting and Gimpel, 2017), followed by long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), and followed finally by word averaging.

The input, whether $\mathbf{x}_{\text{cont}}$ or $\mathbf{x}_{\text{sim+cont}}$, is fed to one fully-connected hidden layer with 300 units, followed by a rectified linear unit (ReLU) activation. Our manually computed features (Length, IDF, PMI, and Mention) are concatenated prior to this layer. The output layer follows and uses a linear activation.

We use mean absolute error as our loss function during training. We train to predict the three criteria jointly, so the loss is actually the sum of mean absolute errors over the three criteria. We found this form of multi-task learning to significantly outperform training separate models for each criterion. When tuning, we tune based on the average Spearman correlation across the three criteria on our validation set. We train all models for 25 epochs using Adam (Kingma and Ba, 2014) with a learning rate of 0.001.

6 Experiments

After averaging the two annotator scores to get our dataset of 4586 context/continuation pairs, we split the data randomly into 600 pairs for validation, 600 for testing, and used the rest (3386) for training. For our evaluation metric, we use Spearman correlation between the scorer's predictions and the annotated scores.

6.1 Feature Ablation

Table 6 shows results as features are either removed from the full set or added to the featureless model, all when using the "cont" input schema. Each row corresponds to one feature ablation or addition, except for the final row which corresponds to adding two feature sets that are efficient to compute: IDF and Length. The Mention and PMI features are the most useful for relevance, which matches the pattern of correlations in Table 5, while IDF and Length features are most helpful for interestingness. All feature sets contribute in predicting overall quality, with the Mention features showing the largest drop in correlation when they are ablated.

	O	R	I
All features	57.3	53.4	49.6
- PMI	56.3	50.4	48.6
- IDF	56.6	53.6	46.0
- Mention	54.8	50.3	48.6
- Length	56.1	55.9	45.3
No features	51.9	44.9	43.8
+ PMI	54.5	50.9	44.9
+ IDF	54.3	46.7	46.3
+ Mention	53.8	48.8	46.0
+ Length	51.9	43.1	44.9
+ IDF, Length	54.6	46.5	47.3

Table 6: Ablation experiments with several feature sets (Spearman correlations on the validation set).

		validation			test		
model	features	O	R	I	O	R	I
cont	none	51.9	44.9	43.8	53.3	46.0	50.5
	IDF, Len.	54.6	46.5	47.3	51.6	40.6	50.2
	all	57.3	53.4	49.6	57.1	54.3	52.8
sim+ cont	none	51.6	43.7	44.3	52.2	45.0	48.4
	IDF, Len.	54.2	45.6	47.7	56.0	46.8	53.0
	all	55.1	54.8	47.4	58.7	55.8	52.9

Table 7: Correlations (Spearman's $\rho \times 100$) on validation and test sets for best models with three feature sets.

6.2 Final Results

Table 7 shows our final results on the validation and test sets. The highest correlations on the test set are achieved by using the sim+cont model with all features. While interestingness can be predicted reasonably well with just IDF and the Length features, the prediction of relevance is improved greatly with the full feature set.

Using our strongest models, we computed the average predicted criterion scores for each story generation system on the test set. Overall, the predicted rankings are strongly correlated with the rankings yielded by the aggregated annotations shown in Table 2, especially in terms of distinguishing human-written and machine-generated continuations.

While the PMI features are very helpful for predicting relevance, they do have demanding space requirements due to the sheer number of word pairs with nonzero counts in large corpora. We attempted to replace the PMI features by similar features based on word embedding similarity, following the argument that skip-gram embeddings with negative sampling form an approximate factorization of a PMI score matrix (Levy and Goldberg, 2014). However, we were unable to find the same performance by doing so; the PMI scores were still superior.

For the automatic scores shown in Table 3, we used the sim+cont model with IDF and Length features. Since this model does not require PMIs or NLP analyzers, it is likely to be the one used most in practice by other researchers within training/tuning settings. We release this trained scorer as well as our annotated data to the research community.

7 Conclusion

We conducted a manual evaluation of neural sequence-to-sequence and retrieval-based story continuation systems along three criteria: overall quality, relevance, and interestingness. We analyzed the annotations and identified features that correlate with each criterion. These annotations also provide a new story understanding task: predicting the quality scores of generated continuations. We took initial steps toward solving this task by developing an automatic scorer that uses features, compositional architectures, and multitask training. Our trained continuation scorer can be used as a rich feature function for story generation or a reward function for systems that use reinforcement learning to learn to generate stories. The annotated data and trained scorer are available at the authors' websites.

Acknowledgments

We thank Melissa Roemmele and the anonymous reviewers for their comments. We also thank NVIDIA Corporation for donating GPUs used in this research.

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*. pages 1171–1179.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 1724–1734.

Angel Daza, Hiram Calvo, and Jesús Figueroa-Nazuno. 2016. Automatic text generation by learning from literary structures. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*. pages 9–19.

Natalie Dehn. 1981. Story generation after TALE-SPIN. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI)*. pages 16–18.

David K. Elson. 2012. *Modeling Narrative Discourse*. Ph.D. thesis, Columbia University.

Albert Gatt and Emiel Krahmer. 2017. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *arXiv preprint arXiv:1703.09902* .

Pablo Gervás. 2012. Story generator algorithms. *The Living Handbook of Narratology* 19.

Andrew S. Gordon, Cosmin Adrian Bejan, and Kenji Sagae. 2011. Commonsense causal reasoning using millions of personal stories. In *25th Conference on Artificial Intelligence (AAAI-11)*.

Andrew S. Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop*.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8).

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, pages 2333–2338.

Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1233–1239.

Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan. 2017. Story generation from sequence of independent short descriptions. In *Workshop on Machine Learning for Creativity, at the 23rd SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2017)*.

Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 329–339.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Sheldon Klein, John F. Aeschlimann, David F. Balsiger, Steven L. Converse, Claudine Court, Mark Foster, Robin Lao, John D. Oakley, and Joel Smith. 1973. Automatic novel writing: A status report. Technical Report 186, University of Wisconsin-Madison.

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

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27*.

Yuhua Li, David McLean, Zuhair A Bandar, James D O'shea, and Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE transactions on knowledge and data engineering* 18(8):1138–1150.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 2122–2132.

Hugo Liu and Push Singh. 2002. MAKEBELIEVE: Using commonsense knowledge to generate stories. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI)*. pages 957–958.

Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Inderjeet Mani. 2012. Computational modeling of narrative. *Synthesis Lectures on Human Language Technologies* 5(3):1–142.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL (System Demonstrations)*. pages 55–60.

Neil McIntyre and Mirella Lapata. 2009. Learning to tell tales: A data-driven approach to story generation. 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 217–225.

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*. pages 1562–1572.

Neil Duncan McIntyre. 2011. *Learning to tell tales: automatic story generation from Corpora*. Ph.D. thesis, The University of Edinburgh.

James R. Meehan. 1977. TALE-SPIN, an interactive program that writes stories. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI)*. pages 91–98.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 839–849.

K.M. Oinonen, Mariet Theune, Antinus Nijholt, and J.R.R. Uijlings. 2006. *Designing a story database for use in automatic story generation*, Springer Verlag, pages 298–301. Lecture Notes in Computer Science.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. pages 311–318.

Vladimir Propp. 1968. *Morphology of the Folktale*, volume 9. University of Texas Press.

Mark O. Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39:217–268.

Melissa Roemmele. 2016. Writing stories with help from recurrent neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*. pages 4311 – 4312.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*. Stanford University.

Melissa Roemmele and Andrew S. Gordon. 2015. Creative help: a story writing assistant. In *International Conference on Interactive Digital Storytelling*. Springer, pages 81–92.

Melissa Roemmele, Andrew S. Gordon, and Reid Swanson. 2017. Evaluating story generation systems using automated linguistic analyses. In *Workshop on Machine Learning for Creativity, at the 23rd SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2017)*.

Roger C. Schank and Robert P. Abelson. 1975. Scripts, plans, and knowledge. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence (IJCAI)*. pages 151–157.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Reid Swanson and Andrew S. Gordon. 2012. Say anything: Using textual case-based reasoning to enable open-domain interactive storytelling. *ACM Trans. Interact. Intell. Syst.* 2(3):16:1–16:35.

Mariët Theune, Sander Faas, Anton Nijholt, and Dirk Heylen. 2003. The virtual storyteller: story creation by intelligent agents. In *Proceedings of the 1st International Conference on Technologies for Interactive Digital Storytelling and Entertainment*. Springer, pages 204–215.

Perry W. Thorndyke. 1977. Cognitive structures in comprehension and memory of narrative discourse. *Cognitive psychology* 9(1):77–110.

Scott R. Turner. 1993. *Minstrel: a computer model of creativity and storytelling*. Ph.D. thesis, University of California at Los Angeles.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424* .

Tong Wang, Ping Chen, and Boyang Li. 2017. Predicting the quality of short narratives from social media. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Charagram: Embedding words and sentences via character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 1504–1515.

John Wieting and Kevin Gimpel. 2017. Revisiting recurrent networks for paraphrastic sentence embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 2078–2088.

Robert Wilensky. 1983. Story grammars versus story points. *Behavioral and Brain Sciences* 6(4):579–591.

Patrick Henry Winston. 2014. The Genesis story understanding and story telling system: A 21st century step toward artificial intelligence. Memo 019, Center for Brains Minds and Machines, MIT.