Evaluating Story Generation Systems Using Automated Linguistic Analyses

Melissa Roemmele, Andrew S. Gordon, and Reid Swanson
Institute for Creative Technologies, University of Southern California
roemmele@ict.usc.edu, gordon@ict.usc.edu, reid@reidswanson.com

ABSTRACT

Story generation is a well-recognized task in computational creativity research, but one that can be difficult to evaluate empirically. It is often inefficient and costly to rely solely on human feedback for judging the quality of generated stories. We address this by examining the use of linguistic analyses for automated evaluation, using metrics from existing work on predicting writing quality. We apply these metrics specifically to story continuation, where a model is given the beginning of a story and generates the next sentence, which is useful for systems that interactively support authors' creativity in writing. We compare sentences generated by different existing models to human-authored ones according to the analyses. The results show some meaningful differences between the models, suggesting that this evaluation approach may be advantageous for future research.

Keywords

story generation, evaluation metrics, linguistic analysis of creativity, assistive creative tools

1. INTRODUCTION

Automated story generation is a long-pursued endeavor in artificial intelligence, and one that been used to propose models of human creativity [12, 46, 56]. This area of research emerged with formal reasoning approaches based on hand-authored rules about narrative structure [29, 32, 38], and has now evolved to accommodate more data-driven methods where knowledge is acquired automatically from story corpora [33, 36, 52]. These data-driven systems have the advantage of promoting open-domain generation, meaning that they can model the diversity of narrative content contained in larger corpora. This is particularly useful for systems that are intended to support the creativity of human authors by interactively generating from author-provided text, where models must be flexible enough to accommodate different story genres.

One of the difficulties of story generation is how to evaluate the quality of the stories empirically, as there are an immense number of 'correct' possibilities for even just the next sentence in a given story. It is common to rely on human judgments of story quality for evaluation; for instance, by asking people to rate stories on different dimensions (e.g. coherence, creativity) [36, 46, 51] or by instructing them to edit the stories, where then the edit distance becomes an inverse measure of quality [33, 48]. However, human evaluations can be time-consuming and costly to carry out, particularly since they must be repeated for each new set of generated stories. While evaluating generation quality in a fully automated way is likely as difficult as the generation task itself, progress on this research would greatly benefit from tools that can provide some indication of generation quality without manual analysis.

For this work, we focus on the task of story generation in a continuation framework, where the system is given multiple sentences of a human-written story and is prompted to generate the subsequent sentence. We use automated linguistic analyses to compare sentences generated by different models to the corresponding sentences in the human-authored stories, which we treat as a gold standard for narrative quality. These metrics involve straightforward natural language processing techniques that have been used to evaluate features of writing quality in existing work, which we detail in Section 4. We apply these analyses to compare existing data-driven approaches: a case-based reasoning approach that uses a nearest-neighbors similarity model to retrieve relevant next sentences; a recurrent neural network approach that predicts sentences word-by-word according to learned probabilities; as well as two relevant random baselines. This work is one of the few to examine the use of automated linguistic analyses in evaluating story generation. In the long-term view, we see this as an initial step in exploring how generation systems can help authors improve their creative writing.

2. STORY CONTINUATION TASK

We apply automated linguistic analyses to evaluate generation systems in a story continuation task. In this task, an initial story is provided as input, and the generation systems output a single next sentence in the story. This design has been used in previous work [48, 51] and is also related to interactive fiction where fixed story segments are selected based on user-provided text (see review in [41]). As discussed in Section 4, analyzing generated content with reference to existing stories can make it easier to quantitatively evaluate, since the features of a generated sentence can be compared to a human-authored gold standard.

We performed this story continuation task on stories from the Children's Book Test[1] [21]. This dataset contains children's novels authored between 1850 and 1950 freely available through Project Gutenberg[2]. Each book is divided into passages of 21 sentences. The intended task for this dataset is to use the first 20 sentences (the context) to predict a word that is missing from the 21st sentence given a set of candidate words. We did not directly attempt this task in this work, but instead used the context of the passage to generate the 21st sentence. We performed generation on only the items in the validation and test sets, which consist of a total of 18,000 passages with 440 average words per context. Table 1 shows two examples of story contexts (in italics) that come from Andrew Lang's *The Grey Fairy Book* and Lucy Maud Montgomery's *The Golden Road*. We used the actual 21st sentence contained in each item as a gold standard with which to compare our models, based on the assumption that this sentence is a high-quality continuation of the story.

The CBT dataset is distinct from others that have been used for story generation research. For example, the ROCStories dataset [43] contains five-sentence stories about stereotypical everyday experiences, where the focus is on predicting what is most likely to happen next based on commonsense expectations. However, these stories are less representative of traditional features of narrative that relate to creativity, such as writing style, character development, and surprise. Since our future goal is to provide creative authoring support, we selected stories known to have these classic characteristics of narrative.

3. GENERATION MODELS

In this work, we evaluated two different models that both take an initial story (context) as input and generate the next sentence: a case-based reasoning (CBR) model and a recurrent neural network (RNN) model. We also considered two baseline methods to further inform our interpretation of the analyses. Table 1 shows examples of sentences produced by each model.

3.1 Training Data

For training the models, we used a different dataset from the CBT stories. This dataset also consists of fiction stories, but from the domain of fiction-writing websites instead of classic literature. One motivation for this difference is that in interactive generation systems, a user's specific story genre might not be known in advance, so this capacity for domain adaptation can be particularly important. To compile this dataset, we gathered stories from websites including **fictionaut.com**, **ficwad.com**, **wattpad.com**, **writerscafe.org**, and various other sites containing fiction uploaded by authors themselves. These stories cover a wide range of genres related to fantasy, horror, romance, and science fiction; many of them are fan fiction that depict characters and settings from existing works (e.g. *Harry Potter*, *Naruto*, and *Twilight*). This dataset consists of 607,627 stories, with a total of 41,458,210 sentences (an average of 68 per story) and 467,023,696 words (an average of 787 per story). For all models, we established a vocabulary of words that occurred at least 25 times in this corpus, which ultimately included 83,292 words. All other words were ignored by the models during training (in the case of the RNN and 1-gram baseline, they were all mapped to a single 'unknown word' token).

3.2 Case-based Reasoning (CBR)

Case-based reasoning is a general AI problem-solving approach where a new problem is solved by consulting a known solution for an existing problem [1]. In the context of story generation, CBR is used to establish an analogy between a new story and an existing story so that the existing story can inform the generation of the new story [55]. [51] applied this paradigm to produce new sentences in a story by retrieving them from a corpus (the 'case library'). In this system, given the most recent sentence in a new story, the system finds the existing sentence in the corpus that is most similar to the new sentence. It then looks at the story in which the existing sentence appears and retrieves the sentence that immediately follows it. The idea is that because of the similarity between the new sentence and the existing sentence, what appears after the existing sentence in its story is also a reasonable prediction for what happens next in the new story. To compute similarity, each sentence is encoded as a bag-of-words vector, whose values are the number of times each word in the vocabulary occurs in that sentence. Then the similarity between two sentences is equal to their vector cosine similarity [34].

We reproduced this same approach here, where the training set described above is used as the case library from which sentences are retrieved. In particular, we segmented the training stories into individual sentences[3]. We then built a similarity index[4] that efficiently retrieves sentences from this corpus based on bag-of-words cosine similarity. To generate the next sentence for a given story context, the model uses this index to find the sentence most similar to the last one in the context, locates the story that this sentence appears in, and then retrieves the sentence that follows it in the story.

3.3 Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) are a general framework for modeling sequence data [13], and are now frequently used in language generation tasks. One approach is to use the RNN as a language model that predicts the probability of word sequences. When trained on a corpus of stories, the RNN learns a conditional probability distribution of each word occurring in a story given the words that precede it. This distribution is computed through a set of nonlinear functions (a hidden layer) that maintain a representation of the story up to a word at a particular timepoint in the sequence. Hidden layers can be stacked so that the output of one is the input of another. We use hidden layers with Gated Recurrent Units [8] (GRUs). The input to the first hidden layer is a word sequence where each word is encoded as a vector of real values (a word embedding). The output of the uppermost hidden layer is passed to a top (softmax) prediction layer which gives the probability distribution of all possible next words in the sequence. Training occurs by minimizing cross-entropy loss such that parameters of the model are optimized to increase the predicted

[1] fb.ai/babi/

[2] gutenberg.org/

[3] Most of the NLP processing in this work, including sentence and word segmentation, part-of-speech tagging, and syntactic chunking, was done using spaCy: spacy.io/

[4] Implemented with gensim: radimrehurek.com/gensim/

	Papa,’ she said, ‘it is not artificial, it is REAL!’ ‘Ugh!’ said all the ladies-in-waiting, ‘it is real!’ ‘Let us see first what is in the other casket before we begin to be angry,’ thought the Emperor, and there came out the nightingale. It sang so beautifully that one could scarcely utter a cross word against it. ‘Superbe! charmant!’ said the ladies-in-waiting, for they all chattered French, each one worse than the other. ‘How much the bird reminds me of the musical snuff-box of the late Empress!’ said an old courtier. ‘Ah, yes, it is the same tone, the same execution!’ ‘Yes,’ said the Emperor; and then he wept like a little child. ‘I hope that this, at least, is not real?’ asked the Princess. ‘Yes, it is a real bird,’ said those who had brought it. ‘Then let the bird fly away,’ said the Princess; and she would not on any account allow the Prince to come. But he was nothing daunted. He painted his face brown and black, drew his cap well over his face, and knocked at the door. ‘Good-day, Emperor,’ he said. ‘Can I get a place here as servant in the castle?’
R-sent	The music plays, it’s my favourite song, But I don’t want to sing along.
CBR	A thousand flashbacks appeared in my mind vision as I demanded my brain to find a reasonable explanation for my present state.
1-gram	It and placed peered of why 7:40 and bolted you.
RNN	Yes, he’s a ‘one’.
Gold	‘Yes,’ said the Emperor, ‘but there are so many who ask for a place that I don’t know whether there will be one for you; but, still, I will think of you.
	“Cecily, you’ve got a dreadful cold,” said the Story Girl anxiously. “In spite of Peg’s ginger tea,” added Felix . “Oh, that ginger tea was AWFUL,” exclaimed poor Cecily. “I thought I’d never get it down – it was so hot with ginger – and there was so much of it! But I was so frightened of offending Peg I’d have tried to drink it all if there had been a bucketful. Oh, yes, it ’s very easy for you all to laugh! You didn’t have to drink it.” “We had to eat two meals, though,” said Felicity with a shiver. “And I don’t know when those dishes of hers were washed. I just shut my eyes and took gulps.” “Did you notice the soapy taste in the porridge?” asked the Story Girl. “Oh, there were so many queer tastes about it I didn’t notice one more than another,” answered Felicity wearily. “What bothers me,” remarked Peter absently, “is that skull. Do you suppose Peg really finds things out by it?” “Nonsense! How could she?” scoffed Felix, bold as a lion in daylight.“ She didn’t SAY she did, you know,” I said cautiously. “Well, we’ll know in time if the things she said were going to happen do,” mused Peter.
R-sent	How in the hell am I supposed to say no to that face?
CBR	I sighed and stood
1-gram	You!
RNN	“Course” I said, then nodded.
Gold	“Do you suppose your father is really coming home?”

Table 1: Examples of story contexts and corresponding next sentences generated by each model

probabilities of the true words that actually appear a story. This model is now very common in NLP; see existing work for its details [9, 39, 49]. After training, the learned distributions can then be used to generate new words in a story by iteratively sampling from the probabilities of all potential next words in the story [50].

In this work, we used an RNN with a 300-dimension embedding layer and two 500-dimension GRU layers[5]. We tokenized the stories into lowercased words, and all punctuation was treated in the same way as word tokens. During training, the model processed entire stories word by word in batches of 50 stories at a time, using the Adam algorithm [27] for optimization. To use the trained model to generate a new sentence for a given context, we fed the context into the model and sampled a word from the probability distribution for the next word. We appended this word to the story as the beginning of the next sentence, and continued adding words to the sentence until an end-of-sentence punctuation token (‘.’, ‘!’ and ‘?’) was generated. As has been done in existing work, e.g. [26], we used a ‘temperature’ variable to smooth some of the randomness in the probability distribution. Because some of our analyses require us to present sentences as regular strings rather than lists of tokens, we ‘detokenized’ the sentences using some heuristics for punctuation formatting, capitalization, and merging contractions.

The RNN has some theoretical advantages over the CBR model. First, it is a productive model, meaning that it can generate sequences that do not directly appear in its training data. Whereas there are an exponential number of sequences the RNN can produce through all possible combinations of words in the vocabulary, the CBR model is limited to the sentences it has observed in the corpus. In this way, the RNN is arguably a better model of computational creativity, since natural language has this same productivity that leads authors to write completely novel text. Another important advantage of the RNN model is that it considers an entire story when generating the next sentence; in contrast, the CBR model only observes the most recent sentence. Consequently, the RNN has more opportunity to recall events or entities that appeared earlier on in the story. Obviously, authors have long-term memory of what they have written previously in the story, so the RNN also has this advantage as a model of creativity. However, it is important to keep in mind that the CBR model retrieves human-authored sentences, which may be favorable in practical ways over the RNN-generated ones. It is interesting to compare these particular models on the same task because while they are both rely on a data-driven approach, they assemble sequences from different units of generation (sentences versus words) and thus are very different models.

[5] Implemented with Keras: keras.io/

3.4 Baselines

We also considered two baseline models as additional comparisons in our analyses, both of which randomly generate sentences without regard to the story context. The first baseline (R-sent) simply selects a random sentence from the training corpus. The second baseline is a unigram language model (1-gram), which like the RNN generates sentences word by word. Its probability distribution is just the relative frequency of each word in the training corpus, so each word is sampled independently from the previous word. By including these in the analyses, we show the expected performance on the metrics even when there is minimal complexity to the model. However, there is another theoretical purpose for considering these baselines for this task. Randomness has been discussed as a desirable feature of creative systems [5, 53, 42]. These baselines may produce sentences that are unusual and surprising given the context, which may be appealing to authors if they can come up with a coherent interpretation of this randomness. While the metrics considered here focus generally on writing quality rather than the effect of randomness on creativity, it is interesting to keep this mind when interpreting the analyses.

4. AUTOMATED LINGUISTIC ANALYSES

We applied a set of automated linguistic analyses to examine differences in the writing quality of the generated sentences within their story context. Each metric is listed below with an explanation of its relevance to this evaluation task. The metrics can be broadly categorized into two types: 1) metrics that analyze the generated sentence in isolation from its context (Story-Independent), and 2) those that evaluate the sentence with reference to the context (Story-Dependent). Intuitively, the first type of analysis captures how well-written the sentence is by itself, while the second determines how apt the sentence is for that particular story; both dimensions are likely important for creative writing generation.

4.1 Story-Independent Metrics

Sentence Length: Even with its simplicity, sentence length is a feature that can reliably discriminate between text genres, authors, and other characteristics like overall readability [15, 19, 25]. Length is simply the number of words in each generated sentence (Metric 1).

Grammaticality: It is generally accepted that high-quality writing minimizes grammatical mistakes [37], and therefore error detection and correction is an active research area [6, 31]. To judge the grammaticality of generated sentences, we used Language Tool[6] [40], a rule-based system that detects various grammatical as well as spelling errors. Using this system we computed an overall grammaticality score (Metric 2) for each sentence, equal to the proportion of total words in the sentence deemed to be grammatically correct.

Lexical Diversity: High-quality writing has been found to contain a larger set of unique words and phrases, and avoids overly repetitious use of the same phrases [7, 10, 24]. We analyzed the number of unique words (types) in the generated sentences relative to the number of total word occurrences (tokens), known as the type-token ratio (Metric 3). A single type-token ratio was computed for each model from the entire set of sentences generated by that model. Because our models were only aware of words that occurred 25 or more times in the training data, we only counted words in this vocabulary in the ratio (in contrast to the R-sent and CBR models, the RNN and 1-gram models never had the opportunity to generate words not in this vocabulary). We also measured the number of unique trigrams in the same way, by computing the total proportion of unique trigrams to the total number of trigram occurrences in the generated sentences (Metric 4), again only considering trigrams where all tokens were contained in the training vocabulary.

Lexical Frequency: Related to lexical diversity, more advanced writing often contains fewer common words [11, 24, 37]. We measured the average log frequency of the words in each generated sentence (Metric 5), where the frequencies were Good-Turing smoothed counts taken from the 3-billion word Reddit Comment Corpus[7]. To keep this metric consistent with the others where higher scores are hypothesized as more favorable, we report the negative (inverse) log frequency, so that higher numbers indicate lower word frequency.

Syntactic Complexity: Existing research has documented that high-profocency writing tends to be more syntactically complex [4, 37, 47, 57, 58]. We examined this complexity in terms of the number and length of syntactic phrases (often called chunks [54]) in the generated sentences. We counted the total number of noun phrases (Metric 6) and words per noun phrase (Metric 7), and equivalently the number of verb phrases (Metric 8) and words per verb phrase (Metric 9). To account for the effect of sentence length on these measures (i.e. longer sentences may naturally contain more and longer phrases), we divided all measures for each sentence by its total number of words.

4.2 Story-Dependent Metrics

Lexical Cohesion: Clearly, a generated sentence should be coherent with the story in which it occurs. There are various dimensions to coherence, one of which is lexical cohesion [20], by which the words in the generated sentence should be semantically related to the words in the story context. While deep modeling of semantics is an open problem, there are some simple shallow metrics for quantifying lexical similarity between text segments [16, 30]. First, and most simply, we computed Jaccard similarity [23] (Metric 10) to find the overall proportion of overlapping words between the context and generated sentence. We filtered this measure to include only words tagged as adjectives, adverbs, interjections, nouns, pronouns, proper nouns, and verbs (with the exception of pronouns, these are the categories associated with content words). Second, we examined similarity in terms of word embeddings, which represent the meaning of a word as an n-dimensional vector of real values, such that words with similar meanings have similar vectors. We specifically used the GloVe embedding vectors [45] trained on the Common Crawl corpus[8]. We computed the mean cosine similarity of the vectors for all pairs of content words between a generated sentence and its context (Metric 11).

Alternatively, in contrast to computing similarity at the word level, we also looked at similarity between full sentence encodings computed by the skip-thought model [28]. Analogous to word embeddings where the model is trained to predict words from other nearby words, the skip-thought

[6] Code at: github.com/cnap/grammaticality-metrics

[7] spacy.io/docs/api/token

[8] spacy.io/docs/usage/word-vectors-similarities

model learns to represent sentences according to their neighboring sentences. We used 4800-dimension sentence vectors trained on the 11,000 books in the BookCorpus dataset[9] to encode each of the sentences in the context as well as the generated sentence. We then computed the cosine similarity between the mean of the context sentence vectors and the generated sentence (Metric 12).

Style Matching: Previous work has shown that automated analyses of writing style can distinguish successful from unsuccessful writers [18, 44]. Since proficient authors exhibit style consistency across a particular work [17], we similarly posit that generated sentences should match the style of their contexts. We examined style similarity in terms of part-of-speech (POS) categories. First, we computed the similarity in the distributions of word categories between the generated sentence and its story context using the same approach as [22] (Metric 13). Specifically, we compared the number of adverbs, adjectives, conjunctions, determiners, nouns, pronouns, prepositions and punctuation tokens. The similarity for each category can be quantified as $1 - \frac{|cat_{context} - cat_{gensent}|}{cat_{context} + cat_{gensent}}$, where *cat* is the proportion of words of that category in that sequence. For a given context and generated sentence we averaged the similarity scores across all categories to get one overall style matching score. In addition to the category distribution of individual words, we also looked at style similarity in terms of POS trigrams [2] (Metric 14). To do this, we computed the Jaccard similarity between the category trigrams in each generated sentence and those in the corresponding context.

Entity Coreference: Similar to the expectation of lexical cohesion between sentences in a text, a generated sentence should refer to entities that have been previously introduced in the story. Although generated sentences can introduce new entities into the story without necessarily being incoherent, entity coreference has been used in existing work for automatically judging coherence [3, 14]. To get an entity coreference rate, we found the proportion of entities (equivalent to noun phrases) in the generated sentence that coreferred to an entity in the corresponding context (Metric 15). Higher coreference rates indicate more entity coherence between the generated sentence and context.

5. RESULTS

Table 2 shows the mean metric scores across all 18,000 generated sentences for each model compared to the original (gold) sentences contained in the CBT stories. Differences between models were statistically evaluated using two-sample Monte Carlo permutation tests, with significance shown at $p < 0.005$ due to Bonferroni adjustment (there are 10 model comparisons, so the alpha level of 0.05 is adjusted to $0.05/10 = 0.005$).

There are several notable results to highlight in this table. First, the sentences generated by the models were much shorter than the corresponding gold sentences, which may reflect the domain difference between the training corpus and the CBT stories. Overall, the gold sentences most often obtained the highest scores on the metrics, which suggests that these metrics are correlated with writing quality. However, this was not a universal finding. Among the story-independent analyses, the order of the model scores was very mixed. One unexpected result was that the RNN had a higher overall grammaticality score than the gold sentences. Since it is probably not the case that the gold sentences are ungrammatical, it is worth exploring whether there were unique features in the gold sentences that the Language Tool scorer consistently recognized as ungrammatical.

Despite the existing findings that better writing uses a greater variety of words, here the gold sentences had a lower type-token and unique trigram ratio than all models except for the RNN. The random baselines actually demonstrated the highest scores on these measures. This could again be a domain-specific difference between the gold and generated sentences, but the fact that the RNN had even lower lexical diversity than the gold sentences is also an interesting consideration. It is probably less surprisingly that the 1-gram model had the highest unique trigram ratio, since it obeys no constraints on which word combinations qualify as grammatical. There is likely a middle ground for the use of unique phrases that balances creativity with conventionality, since the 1-gram model does not fare better on the other quality measures. It may also be the case that in work where lexical diversity is modeled as a feature of good writing, a comparison was made between more and less proficient human writers, rather than with random methods. On the other hand, the findings for word frequency favored the human-authored sentences as expected, as these sentences did use less frequent words. Along with having a narrower vocabulary, the RNN model tended towards more common words relative to the other models.

In terms of syntactic complexity, the gold sentences appeared to contain far more noun and verb phrases than the other models, but this was not the case once sentence length was taken into account. While it was expected that the 1-gram sentences had little syntactic complexity (since the model has no knowledge of syntax), it was surprising that the gold phrases were also much shorter on average than the phrases generated by the other models. The RNN model was notable for its verb phrases, which were longer and more frequent than those in the other sentences.

It is intriguing that significant differences emerged between the R-sent and CBR models on the story-independent metrics, since these sentences come from the same corpus and therefore would be expected to have similar features. It is likely that the CBR model selected sentences with specific features not evenly distributed across the corpus at large; for example, the CBR sentences were longer, more grammatical, contained rarer words, and had less verb phrase complexity.

The results for the story-dependent analyses are more consistent across metrics. For all measures, the gold sentences scored the highest: they were more semantically related to their story contexts, were more stylistically similar in terms of part-of-speech categories, and the entities they mentioned were more likely to corefer to those in the context. This, along with the low performance of the random baselines on these measures, suggests that these metrics do capture story coherence.

The story-dependent metrics are particularly useful for comparing the CBR and RNN models. In Section 3, we discussed how in contrast to the CBR model, the RNN can in theory observe the entire story context and produce sentences more targeted to that unique context. We observe some practical evidence for this in these results: the RNN sentences were more semantically related to the context in terms of Jaccard and skip-thought similarity, and

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

	R-sent	**CBR**	**1-gram**	**RNN**	**Gold**
Story-independent metrics					
1. Sentence length	13.36	15.56*‡§	13.67	13.10	28.84*†‡§
2. Grammaticality	0.957‡	0.961*‡	0.925	0.992*†‡⋆	0.982*†‡
3. Type-token ratio	0.057†§⋆	0.042§⋆	0.057†§⋆	0.010	0.020§
4. Unique trigram ratio	0.776†§⋆	0.491§⋆	0.946*†§⋆	0.307	0.418§
5. Inverse word frequency	7.078§	7.122*‡§	7.038§	6.056	7.399*†‡§
6. Noun phrases	0.238‡§	0.239‡§	0.192	0.227‡	0.225‡
7. Noun phrase length	0.149⋆	0.141⋆	0.144⋆	0.143⋆	0.087
8. Verb phrases	0.190†‡⋆	0.186‡⋆	0.164	0.191†‡⋆	0.181‡
9. Verb phrase length	0.367†‡⋆	0.346‡⋆	0.261⋆	0.403*†‡⋆	0.219
Story-dependent metrics					
10. Jaccard similarity	0.004‡	0.005*‡	0.003	0.006*†‡	0.036*†‡§
11. GloVe similarity	0.227‡	0.228‡	0.192	0.227‡	0.246*†‡§
12. Skip-thought similarity	0.682	0.713*	0.718*†	0.733*†‡	0.799*†‡§
13. Word POS similarity	0.503‡	0.541*‡§	0.442	0.501‡	0.698*†‡§
14. Trigram POS similarity	0.028‡	0.034*‡§	0.016	0.031*‡	0.070*†‡§
15. Entity coreference rate	0.440‡	0.456*‡	0.328	0.536*†‡	0.644*†‡§

Statistical significance, $p < 0.005$: *greater than R-sent; †greater than CBR; ‡greater than Unigram; §greater than RNN; ⋆greater than Gold

Table 2: Mean scores on metrics for sentences generated by each model and gold sentences

also more frequently referred to entities introduced in the context. However, the CBR sentences still demonstrated greater stylistic similarity to their contexts than the RNN sentences.

6. DISCUSSION

Overall our results suggest that automated linguistic analyses can capture meaningful differences between generation models, particularly in a story-continuation framework where generated sequences can be directly compared between models. Given the results, particularly for the story-independent analyses, it may not necessarily be the goal of a system to maximize scores on these metrics. Instead, the gold standard sentences help interpret the comparison between the models. If the goal is to make the generated sentences more like the human-authored ones, then progress can be evaluated in terms of the similarity between the gold sentences and the generated ones.

We made use of existing NLP techniques to measure generation quality according to existing indicators of good writing. Because automatically evaluating writing quality is an ongoing research problem, we acknowledge that the analyses used here are relatively shallow metrics based on the tools available. There are more advanced linguistic analyses that we plan to explore in the future, such as entity grids [3] and discourse parsing [35], which may detect deeper coherence relations in stories.

There is an open question of exactly what type of linguistic analyses focus specifically on creativity. Our metrics do not directly address some of the more complex dimensions specific to the domain of narrative, such as character development, plot structure, and suspense. Automatically modeling these type of features is an extremely difficult language understanding problem, for which developing effective metrics still requires much research. However, keeping in mind our goal of providing automated creative assistance, we see any characteristics that make generated content more appealing to authors as relevant to creativity. This work proposes the exploration of automated linguistic analyses for identifying these characteristics.

7. ACKNOWLEDGMENTS

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.

8. REFERENCES

[1] A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 7(1):39–59, 1994.

[2] S. Argamon, M. Koppel, and G. Avneri. Routing documents according to style. In *First International Workshop on Innovative Information Systems*, pages 85–92, 1998.

[3] R. Barzilay and M. Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, Mar. 2008.

[4] S. F. Beers and W. E. Nagy. Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing*, 22(2):185–200, 2009.

[5] M. A. Boden. *The Creative Mind: Myths and Mechanisms*. Basic Books, Inc., New York, NY, USA, 1991.

[6] J. Burstein, M. Chodorow, and C. Leacock. Automated essay evaluation: The Criterion online writing service. *AI Mag.*, 25(3):27–36, Sept. 2004.

[7] J. Burstein and M. Wolska. Toward evaluation of writing style: Finding overly repetitive word use in student essays. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*, pages 35–42, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[8] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation*, page 103, 2014.

[9] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In *Deep Learning and Representation Learning Workshop*, 2014.

[10] S. Crossley, Z. Cai, and D. S. McNamara. Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In *Twenty-Fifth International FLAIRS Conference*, 2012.

[11] S. A. Crossley, J. L. Weston, S. T. M. Sullivan, and D. S. McNamara. The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, 28(3):282–311, 2011.

[12] N. Dehn. Story Generation After TALE-SPIN. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 16–18, 1981.

[13] J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[14] M. Elsner and E. Charniak. Coreference-inspired coherence modeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 41–44, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[15] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221, 1948.

[16] P. Foltz, W. Kintsch, and T. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2&3):285–307, 1998.

[17] M. Gamon. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 611–617, Geneva, Switzerland, Aug 23–Aug 27 2004. Association for Computational Linguistics.

[18] V. Ganjigunte Ashok, S. Feng, and Y. Choi. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1764, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[19] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2):193–202, 2004.

[20] M. A. K. Halliday and R. Hasan. *Cohesion in English*. Longman, London, 1976.

[21] F. Hill, A. Bordes, S. Chopra, and J. Weston. The goldilocks principle: Reading children's books with explicit memory representations. In *International Conference on Learning Representations*, 2016.

[22] M. E. Ireland and J. W. Pennebaker. Language style matching in writing: synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, 99(3):549, 2010.

[23] P. Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, Feb. 1912.

[24] J. Kao and D. Jurafsky. A computational analysis of style, affect, and imagery in contemporary poetry. In *NAACL Workshop on Computational Linguistics for Literature*, pages 8–17, 2012.

[25] J. Karlgren and D. Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 1071–1075, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[26] C. Kiddon, L. Zettlemoyer, and Y. Choi. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339, Austin, Texas, November 2016. Association for Computational Linguistics.

[27] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations*, San Diego, May 2015.

[28] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 3294–3302, Cambridge, MA, USA, 2015. MIT Press.

[29] S. Klein, J. Aeschlimann, and D. Balsiger. Automatic novel writing: A status report. *Wisconsin University*, 1973.

[30] M. Lapata and R. Barzilay. Automatic evaluation of text coherence: Models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1085–1090, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[31] C. Leacock, M. Chodorow, M. Gamon, and J. Tetreault. Automated grammatical error detection for language learners. *Synthesis Lectures on Human Language Technologies*, 7(1):1–170, 2014.

[32] M. Lebowitz. Story-telling as planning and learning. *Poetics*, 14(6):483–502, 1985.

[33] B. Li, S. Lee-Urban, G. Johnston, and M. Riedl. Story Generation with Crowdsourced Plot Graphs. In *27^{th} AAAI Conference on Artificial Intelligence*, 2013.

[34] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[35] D. Marcu and A. Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[36] N. McIntyre and M. Lapata. Learning to tell tales: A data-driven approach to story generation. In *Proceedings of the Joint Conference of the 47^{th} Annual Meeting of the ACL and the 4^{th} International Joint Conference on Natural Language Processing of the AFNLP*, pages 217–225, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[37] D. S. McNamara, S. A. Crossley, and P. M. McCarthy. Linguistic features of writing quality. *Written Communication*, 27(1):57–86, 2010.

[38] J. R. Meehan. TALE-SPIN, An Interactive Program that Writes Stories. In *5^{th} International Joint Conference on Artificial Intelligence*, pages 91–98, 1977.

[39] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and

S. Khudanpur. Recurrent Neural Network based Language Model. In *Proceedings of the 11^{th} Annual Conference of the International Speech Communication Association*, pages 1045–1048, 2010.

[40] M. Miłkowski. Developing an open-source, rule-based proofreading tool. *Softw. Pract. Exper.*, 40(7):543–566, June 2010.

[41] N. Montfort. *Twisty Little Passages: an approach to interactive fiction*. MIT Press, 2005.

[42] N. Montfort and N. Fedorova. Small-scale systems and computational creativity. In *International Conference on Computational Creativity*, page 82, 2012.

[43] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, 2016.

[44] J. W. Pennebaker, C. K. Chung, J. Frazee, G. M. Lavergne, and D. I. Beaver. When small words foretell academic success: The case of college admissions essays. *PLOS ONE*, 9(12):1–10, 12 2015.

[45] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[46] R. Pérez y Pérez and M. Sharples. MEXICA: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence*, 13(2):119–139, 2001.

[47] E. Pitler and A. Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[48] M. Roemmele and A. S. Gordon. Creative help: A story writing assistant. In *International Conference on Interactive Digital Storytelling*. Springer International Publishing, 2015.

[49] M. Sundermeyer, R. Schlüter, and H. Ney. Lstm neural networks for language modeling. In *INTERSPEECH*, pages 194–197, 2012.

[50] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28^{th} International Conference on Machine Learning*, pages 1017–1024, 2011.

[51] R. Swanson and A. S. Gordon. A Comparison of Retrieval Models for Open Domain Story Generation. *Artificial Intelligence*, pages 119–126, 2009.

[52] R. Swanson and A. S. Gordon. Say Anything: Using Textual Case-Based Reasoning to Enable Open-Domain Interactive Storytelling. *ACM Transactions on Interactive Intelligent Systems*, 2(3):1–35, 2012.

[53] J. Sweller. Cognitive bases of human creativity. *Educational Psychology Review*, 21(1):11–19, 2009.

[54] E. F. Tjong Kim Sang and S. Buchholz. Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning-Volume 7*, pages 127–132. Association for Computational Linguistics, 2000.

[55] S. R. Turner. A case-based model of creativity. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, 1991.

[56] S. R. Turner. *MINSTREL: A computer model of creativity and storytelling*. University of California at Los Angeles, 1993.

[57] E. von Glasersfeld. The problem of syntactic complexity in reading and readability. *Journal of Literacy Research*, 3(2):1–14, 1970.

[58] W. Yang, X. Lu, and S. C. Weigle. Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28:53 – 67, 2015.