Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Novel Approach for Effective Automatic Story Generation in English Language

Jainin Sanjaykumar Vakil[1*], Nirbhay Kumar Chaubey[2]

[1*]Faculty of Computer Applications, Ganpat University, Gujarat, India, [1]jsvakil94@gmail.com orchid id: (0009-0001-1297-5177)
[2]Faculty of Computer Applications, Ganpat University, Gujarat, India orchid id: (0000-0001-6575-7723)

ARTICLE INFO

Received: 05 Nov 2024
Revised: 17 Dec 2024
Accepted: 27 Dec 2024

ABSTRACT

This study investigates the connections among different automated evaluation criteria for the creation of short stories. The N-gram model, the CBOW (Continuous Bag-of-Words) model, the GRU (Gated Recurrent Unit) model, and the Generative Pre-trained Transformer 2 (GPT-2) model are among the language models it uses to generate texts from short stories. Aesop's brief stories are used to instruct all models. The produced texts are assessed using a number of metrics, such as WMD (Word Mover's Distance), BERTScore, Perplexity, BLEU score, the quantity of grammatical errors, Self-BLEU score, and ROUGE score. When these evaluation measures are correlated, four different clusters of metrics with significant relationships are found. The first cluster shows a moderate correlation between perplexity and grammatical errors. The second group reveals a strong correlation between BLEU, ROUGE, and BERTScore. In contrast, WMD exhibits a negative correlation with BLEU, ROUGE and BERTScore. Furthermore, Self-BLEU, which measures the diversity of the generated text, shows no significant correlation with any of the other metrics. Ultimately, the study concludes that a comprehensive evaluation of generated text requires the use of multiple metrics, each focusing on a different characteristic of the text quality.

Keywords: NLP, Rough, Story, Generation, English

1. Introduction

A key aspect of human communication is storytelling. Stories serve as a powerful means of connecting with others. When narratives are shared effectively, people become more engaged and absorb more information from them. However, computers still have a long way to go in mastering the art of storytelling. Storytelling enhances communication between humans and machines, and advancements in natural language processing are driven by the development of automated storytelling. Research in computational storytelling encompasses understanding, representing, and creating narratives [1][2]. A narrative is a sequential account of events that have occurred, shaped by the relationships (whether friendly, antagonistic, or romantic) among the characters. People's perspectives and decisions regarding the events in their lives are expressed through stories, which also provide enjoyment and information to others. A story is made up of particular events (such as narration and chronology), a theme that emerges from related occurrences, and a plot that weaves the entire narrative together. By exploring the connections between events and other occurrences, we can uncover relationships and new insights. In narrative forms such as diaries or autobiographies, stories can sometimes be used to meet personal goals.It is also employed in various areas such as social media, user experience design, and marketing to effectively communicate important information [3][4].
One of the most creative activities that helps individuals move from being readers to writers is storytelling. The techniques for crafting short stories have changed dramatically with the introduction of sophisticated natural language generation systems such as GPT-2, BART, and others. A significant challenge with automated story generation is the difficulty in maintaining coherence throughout the narrative. The best way to ensure coherence is to plan each paragraph in advance, similar to how one would approach writing a novel. To maintain coherence in storytelling, creators carefully select the elements that make up the narrative, including characters, themes, and settings [5][6]. One of the most important aspects of automated narrative generation is ensuring consistency at the paragraph level, as a story is made up of various scenes, each containing multiple paragraphs. In the same way that humans design a tale before writing, the system must (1) create a storyline and (2) produce paragraphs that fit the plot. The system may struggle to create a

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

narrative with a smooth flow if it tries to generate the following paragraph without any prior planning. Various studies introduce planning techniques to maintain a story's cohesion [7][8]. These techniques include aligning scene-level circumstances with events and leveraging character traits. Other methods utilize common sense and global planning to identify key concepts. These methods, however, do not offer ongoing system monitoring; rather, they concentrate on creating each paragraph simultaneously rather than sequentially. The method requires a controller that can provide the correct instructions to create a narrative that effectively captures the intended flow.

Recent advancements in technology have sparked a growing interest in computational narrative, or the automated generation of stories [9][10]. This field is important because it enhances the interaction between humans and intelligent systems. Key elements of Automatic Story Generation (ASG) include computational creativity and artificial intelligence (AI). ASG focuses on using algorithms to craft narratives. The primary techniques for generating stories automatically include neural network-based methods, rule-based approaches, planning-based strategies, Case-Based Reasoning (CBR), stochastic and probabilistic methods, transformer models, and interactive storytelling. Early approaches to automated story generation (ASG) utilized rule-based systems that depended on predefined rules and templates. These deterministic systems were rooted in logic theory and syntax patterns. One notable example is the “Tale-Spin” system from the 1970s, which simulated characters with specific goals and plans to create simple narratives. While these stories were coherent, they often lacked depth. In planning-based systems, the narrative was treated as a planning problem, where the system devised a sequence of actions to reach a narrative goal. These systems effectively illustrated the characters' actions and interactions, contributing to a coherent plot.The “Plot Machines" [11][12] framework focuses on employing AI planning techniques to generate stories. It begins with the system identifying the sequence of events needed to achieve a specific outcome.

Case-Based Reasoning (CBR) approaches create stories by adapting and reusing elements from existing narratives. The system retrieves a story from a database that closely resembles the desired one and modifies it to meet new requirements. An example of this is the “Mexica” system, which generates stories by balancing story tension and coherence, using existing stories as references. Stochastic and Probabilistic models employ probability methods such as Markov chains, HMMs (Hidden Markov Models), or probabilistic context-free grammars to generate stories. These methods characterize the likelihood of certain words or sequences of events [13][14]. Markov models have been used to compose texts in a way that uses the previous word or phrase to forecast the next one, resulting in statistically coherent stories, although they can sometimes be quite disjointed. The latest advancements in Automated Story Generation (ASG) are primarily driven by neural deep learning models. The models analyse various patterns from extensive datasets of narratives, allowing them to create text that resembles human writing. Initially, older neural methods like RNNs and their variants, such as LSTMs, were employed to address the sequential aspects of storytelling. However, they often struggled with maintaining coherence over longer passages. The introduction of transformers, like GPT (Generative Pre-trained Transformer), has significantly improved automatic story generation. These models predict the next word in a sentence and excel at crafting stories that are generally logical and coherent, even in lengthy formats. GPT-3 and GPT-4 are prominent transformer-based models known for their ability to generate intricate and cohesive narratives. Hybrid approaches involve combining different methods to leverage the strengths of each.Combining rule-based systems with neural networks can create a story that adheres to a specific logical structure while also being fluent. These systems utilize planning for plot generation alongside neural networks for natural language processing (NLP), which helps improve the narrative's flow [15][16]. Some of the ASG systems focus on interactive storytelling, allowing users to influence the direction of the story. These systems constantly seek user input, resulting in a narrative that is a coherent blend of planning and machine learning techniques. This approach is evident in interactive fiction games or AI-driven narrative experiences, where the plot evolves based on player choices.

2. Literature Review

P. Li, et al. (2024) suggested a multi-granularity feature fusion (MGF) framework aimed at generating endings for image-guided stories [17]. To grasp the sentiment aspects of the image as part of the overall features, they first made use of an image sentiment extractor. Next, they developed a scene subgraph picker that selected the most pertinent scene subgraph to obtain the image traits of the important area. Finally, they integrated the visual and textual components from the global, region, and object levels. Their algorithm effectively captured the main region features and visual sentiment of the imagery, leading to a more coherent and empathetic conclusion. The findings of experimentation indicated that the MGF framework surpassed the latest versions across most performance indicators.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

G. Wu, et al. (2024) presented a novel approach to creating data stories that customized the narrative using adaptive machine-guided user input [18]. Using a plug-in module made for pre-existing narrative production systems, this approach solicited user involvement via interactive questions derived from the dataset and previous talks. This adaptability improved the system's understanding of user intents and made sure the finished story achieved its goals. By creating an interactive prototype named Socrates, the study demonstrated the efficacy of this novel strategy. Comparing this approach to a top data story production algorithm in a quantitative investigation with eighteen participants, they found that Socrates produced tales that were more relevant and had a higher overlap of insights than stories produced by humans. Additionally, they assessed the usableness of Socrates through discussions with three data analysts and identified potential areas for future development.

T. Rahman, et al. (2023) presented a groundbreaking autoregressive diffusion-dependent system that incorporated a visual memory module. The background and actor context were both successfully captured by this module in the produced frames [19]. By employing sentence-conditioned soft attention on these memories, the framework improved reference resolution and learned to maintain consistency in scenes and characters as per requirement. The researchers added new characters, backdrops, and multi-sentence stories to the MUGEN dataset in order to demonstrate the efficacy of this methodology. The MUGEN, PororoSV, and FlintstonesSV datasets were used in story generation experiments, which showed that this approach not only produced visually appealing frames that complemented the story but also created suitable correspondences between characters and their backgrounds.

Y. Xie, et al. (2022) sought to develop tale endings that were more in line with the situation by using contrastive learning. There were two main challenges in using different learning methods for Story Ending Generation (SEG). The first challenge involved effectively sampling negative endings that do not align with the story context [20]. To address these challenges, they proposed a novel framework called CLseg (Contrastive Learning for Story Ending Generation), this comprised two essential steps: story-specific contrastive learning and multi-aspect sampling. Regarding the initial challenge, they employed an innovative multi-aspect sampling technique to generate incorrect story endings by considering order consistency, causality, and sentiment. To tackle the second challenge, they designed a story-specific contrastive training strategy tailored for story ending generation. Experimentation revealed that CLseg outclassed standard methods and produced story endings with improved uniformity and logic.

A. Raza Samar, et al. (2022) investigated the process of story generation using user-defined contexts or prompts [21]. They introduced a narrative generation architecture called NGen-Transformer, which is based on GP2. This architecture specifically emphasized the context provided by users to create expressive stories. To evaluate their model, they utilized the WritingPrompts dataset, which contained a substantial number of manually written sample stories linked to various prompts or titles. The tests indicated that the NGen-Transformer surpassed many sequence-to-sequence and attention-based models in generating stories.

G. Chen, et al. (2021) presented a neural story generation technique for explainable plot creation with the goal of generating narratives that are fluid, logical, and comprehensible [22]. Unlike traditional approaches, this model could automatically learn to create a high-level plot that links the title to the story. Tests conducted on two standard datasets demonstrated that this method surpassed current leading techniques in neural story generation, as evidenced by both automatic and human evaluations. In this study, they concentrated on bridging the gap between a title and a short story with a one-sentence outline. But simulating the interdependencies between sentences in longer narratives continued to pose a significant challenge. Developing more effective strategies to enhance coherence at the story level is crucial.

D. Shi, et al. (2021) presented Calliope, an innovative system for generating visual data stories from input spreadsheets through an automated process. This system also allowed for easy revisions of the generated stories using an online story editor [23]. Interestingly, this system used a new logic-oriented Monte Carlo tree search method that gradually created tale elements (i.e., data facts) and arranged them logically by navigating the data space supplied by the input spreadsheet. Information theory was used to evaluate the significance of these data facts, and each fact was graphically depicted in a chart with a description generated by a machine. The effectiveness of this approach was estimated through three example stories, two precise tests, and interviews with 10 functional specialists. The results indicated that Calliope significantly enhanced the efficiency of visual data story generation.

J. -W. Lin, et al. (2020) introduced a Syntax-Guided Machine Reading Comprehension (SG-Net) framework for generating stories [24]. This framework utilized Chinese word vectors and learned from Chinese datasets. They also developed a SG-GAN (semi-supervised self-growing generative adversarial network) to produce more truthful sequences. The researchers created a set of tests that altered the input sequences' semantic content in order to evaluate the machine's text quality. According to the experimental findings, SG-Net and

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

SG-GAN were both able to write coherent articles by understanding fundamental grammar and semantics. In conclusion, SG-Net outperformed SG-GAN in understanding more sophisticated semantics and grammar in addition to recalling previously read sentences.

L. Wang, S. Qin, et al. (2019) introduced a novel method for generating translation projects in neural story generation, with the goal of crafting smooth, coherent, and believable narratives [25]. This model distinguishes itself from existing systems by effectively managing high-level processes, such as linking nouns to the storyline and integrating diverse data sources. Testing on two datasets has shown that this method surpasses current state-of-the-art systems in both automated and human evaluations. While they successfully utilized descriptive sentences to connect titles and short stories, modeling sentences in longer narratives remains a challenge. To address this, they plan to enhance their approach by generating multi-sentence descriptions to ensure a unified and compelling narrative.

3. Research Methodology

This study produced texts using various language models that were all trained in the same field. A short narrative domain was chosen for this purpose. The same corpus, but with different language models, was employed to conduct the studies. Once each model was trained, new text was generated from the same starting material. More details about the language model and the corpus used are provided in the next subsection. The investigational technique of the study is depicted in Figure 1. Initially, the corpus is pre-processed into targets and inputs. After training, new texts are generated using the language models. The generated texts are then evaluated based on a variety of criteria.

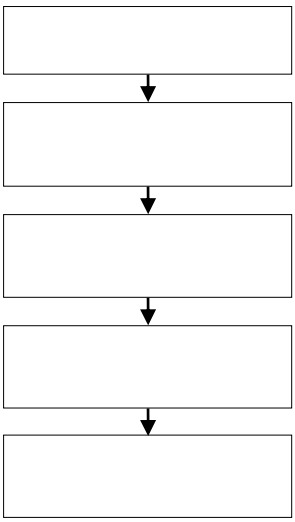

Figure 1: Pipeline of the experiments for Text Generation and Evaluation.

All the steps depicted in the above figure are discussed below:
i. Short story collection: This work selects short stories for the first stage of the experimental process from the Aesop's Fables collection available on americanliterature.com. These fables consist of popular tales. For this study, this work trains language models using 160 short stories from Aesop's Fables. There are 160 stories in the complete corpus, with 32,031 words and 3,418 distinct terms. The stories range in length from 74 to 520 words, with an average of about 186 words per story. The corpus's lexical variety is 10.67%; a greater percentage denotes a more extensive vocabulary. Each story's lexical diversity falls between 37.58% to 75.67%. Each narrative showcases a broader range of terms, as the lexical diversity of individual stories exceeds that of the overall corpus. However, when considering all the stories collectively, the language of the corpus does not show significant variation overall.
ii. Training Corpus Preparation: The corpus is designed to be compatible with language models before the training process begins. Predicting the following word based on the stories' context is the output. The corpus's stories are first transformed into word tokens. Each story is assigned special tokens: an “end-of-story” token is placed after the last word, and a “begin-of-story” token is added before the first word. When training a non-transformer model, It needs a certain number of input nodes or a predetermined amount of context words for the input nodes. Therefore, when K represents the desired number of context words, this

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

sliding window approach will capture each K word. The input consists of a sequence of words, $W_i, W_{i+1}, W_{i+2}, \ldots\ldots W_{i+k}$, and the goal is to predict the next word in this sequence, which is W_{i+k+1}. This process is repeated for each word, moving one word at a time until reaching the end of the sequence. The method is a little different, though, when getting texts ready for transformer model training. Limiting the input to a predetermined word count in the context is not necessary. The initial word in the text establishes the context, and the next word in that context is the output.

iii. Language Model Training: Five different kinds of language models are used in this work. The statistical language modeling method is an N-gram model. A neural network model called a CBOW (Continuous Bag-of-Words) model is used. A recurrent neural network model is used to create a GRU model. Furthermore, two transformer models from Generative Pretrained Transformer 2 (GPT-2) are employed. The pre-trained model (Pretrained GPT-2) and transfer learning (Finetuned GPT-2) were chosen over training a model from scratch due to the GPT-2 model's size and several hyperparameters. These models were selected not only because of their different architectures but also because of the times in which each model became well-known in the language modeling community. Since its inception, the N-gram model has been a frequently used and fundamental approach in language modeling. By introducing embeddings, the CBOW model improves language models' ability to understand meaning. The GRU model, with its ability to capture word sequences through recurrent neural networks, represents a significant advancement over traditional RNNs.The Transformer architecture represents a significant leap forward in language modeling. It is trained on a vast language model, leading to impressive text generation capabilities. When GPT-2 was released, it marked a pivotal moment in the field. Its influence spans numerous sectors, enabling it to produce text that closely resembles human writing in various contexts. This makes GPT-2 one of the most crucial and influential models in the landscape. There are several models available, each with its own unique history regarding its development and rise in the language modelling field. Consequently, the models selected for this study are highlighted. Furthermore, this research evaluates the performance of each model in the chosen areas.

iv. Short Story Generation: The process of generating text with a language model occurs after it has undergone training. It utilizes a left-to-right prediction method to create text. The model adds the term with the highest probability to the resulting text at each stage. The model selects the next word based on the probability weights of possible predictions in order to increase the output's diversity. In essence, a word's probability increases with the likelihood that it would be chosen as the subsequent output. Since the input length is fixed in non-transformer models, the text generation method entails relocating the input window by deleting the first word and adding the next. The input length for transformers does not need adjustment since the model-embedded padding capability for GPT-2 is limited to 1,024. Algorithm 1 presents the methodology.

Algorithm 1 Text Generation Algorithm for Transformer Model

1: **Procedure** $GenerateText(model, startingText, k, maxLength)$
2: $input \leftarrow lastkwordsofstartingText$
3: $genText \leftarrow startingText$
4: $whilelength(genText) \leq maxLength \wedge END_{TOKEN}notfound$
do
5: $nextWordProbs \leftarrow model.probs(input)$
6: $nextWord \leftarrow RandomlySelectWord(nextWordProbs)$
7: $genText \leftarrow genText + nextWord$
8: $input \leftarrow input + nextWord$
9: $returngenText$
10: $EndProcedure$

The first sentence from Aesop's Fables serves as the beginning point for each model in this study. To evaluate and compare the text generated by different models, this research focuses on three key factors: how closely the text resembles human writing (using a corpus), the quality of grammar (measured by the number of grammatical errors), and the diversity of text produced by the same model. Consequently, three stories are generated from each initial sentence.

v. Evaluation: To evaluate the model's performance, the results must be examined. For a variety of reasons, seven distinct scores were chosen. Because of their popularity and the quantity of grammatical errors utilized to assess rule-based grammar, this work uses the Perplexity and BLEU scores. When calculating the distance from the source text, Word Mover's Distance, BERTScore evaluates the deployment capability of the transformer, ROUGE-L score captures the longest common word sequence, and Self-BLEU measures variance within the same model. A brief description of each of these measures can be found below.:

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Perplexity: These metric measures how unpredictable the generated text is. A lower value indicates more human-like and fluent text, while a higher value suggests poor coherence.
Number of Grammatical Errors: This metric rule-based detection (LanguageTools) to count grammatical mistakes, reflecting text quality.
BLEU (Bilingual Evaluation Understudy): This metric compares generated text with reference text using N-gram precision and brevity penalty. A score closer to 1 indicates higher similarity.
Self-BLEU: This metric evaluates text diversity by comparing generated outputs within the same model. A lower score suggests more diverse text.

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation): This metric measures the longest common word sequences between generated and reference text, assessing fluency and coherence.
BERTScore: This metric uses contextual embeddings from the BERT model to compare generated text with reference text, capturing semantic similarities.
Word Mover's Distance (WMD): This metric compute semantic similarity by measuring the minimum word movement needed to transform generated text into reference text. A lower value means higher similarity.

4. Result and Discussion

The evaluation measures' results are shown in this part along with an analysis of their correlation. The metrics for each model are presented in Table 1. For each of the five language models—N-gram, CBOW, GRU, Pretrained GPT-2, and Finetuned GPT-2—it contains the mean, standard deviation, and median for all chosen automatic evaluation measures.

Table 1. Average metric value for each model and evaluation.

Model	Stats	Metrics						
		PPL	Gram Error	BLEU	Self BLEU	ROUGE	BERT	WMD
N-gram	Med	**697.26**	**0.00**	0.119	0.51	0.29	0.55	1.85
	Avg	**703.99**	0.60	0.12	0.52	0.29	0.55	1.87
	SD	101.03	0.74	0.08	0.11	0.07	0.04	0.21
CBOW	Med	1958.97	7.00	0.029	0.49	0.21	0.42	2.07
	Avg	1957.81	8.10	0.029	0.50	0.21	0.42	2.08
	SD	256.24	4.72	0.01	0.08	0.03	0.02	0.15
GRU	Med	850.18	4.00	0.033	**0.37**	0.23	0.45	1.91
	Avg	853.15	4.68	0.036	**0.38**	0.23	0.45	1.91
	SD	117.30	2.99	0.02	0.06	0.04	0.02	0.15
Pretrained GPT2	Med	1141.59	**0.00**	0.149	0.39	0.30	0.54	1.78
	Avg	1184.76	0.62	0.160	0.41	0.31	0.55	1.78
	SD	252.22	0.90	0.07	0.08	0.07	0.04	0.19
Finetuned GPT2	Med	869.71	**0.00**	**0.151**	0.42	**0.30**	**0.58**	**1.70**
	Avg	886.37	**0.54**	**0.158**	0.44	**0.32**	**0.58**	**1.71**
	SD	154.03	0.75	0.08	0.10	0.07	0.04	0.18

Table 1 displays each model's average for each evaluation metric. For example, the average BERTScore of the N-gram model is 0.55, the average Word Mover's Distance is 1.87, the average BLEU score is 0.12, the average Self-BLEU score is 0.52; the average ROUGE-L score is 0.29; the average BERTScore is 0.55; the average perplexity is 703.99; and the average number of grammatical errors is 0.60. A variety of evaluation measures serve as the foundation for the observations on various language models. N-gram, CBOW, Pretrained GPT-2, and Finetuned GPT-2 are the models under comparison. Perplexity, Gram Error, BLEU, Self-BLEU, ROUGE-L, BERTScore, and Word Mover's Distance are the evaluation metrics employed. N-gram has the lowest median and average perplexity, though not by much, when compared to a number of different models. In contrast, CBOW has the greatest perplexity score, meaning it produced the most confusing text. N-gram and Pretrained GPT2 had the lowest grammatical error scores, whereas CBOW has the highest median and average values. While N-gram and Pretrained GPT-2 both performs quite well in this metric, Finetuned GPT-2 has the best BLEU score, indicating superior translation quality or text production. The fact that N-gram has the greatest Self BLEU scores may suggest that the resulting text lacks diversity. The ROUGE metric, which is frequently used to assess the caliber of summaries, is led by the refined GPT-2. The BERTScore measure is another result of it. Last but not least, CBOW has the greatest WMD, suggesting that it may produce language that is more dissimilar from the source. Since the generated text has a semantic meaning that is closer to the reference, lower WMD is typically preferred. Table 7 illustrates how well the Finetuned GPT-2 model fared in practically every category. The model outperforms the best-performing

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

model on the Self-BLEU metric and scores highly on the BLEU, ROUGE, and BERTScore evaluations. It is also among the models that produce text with the fewest grammatical errors.

TABLE 2. Correlation Relation between automatic metrics.

	PPL	Gram. Errors	BLEU	Self BLEU	ROUGE	BERT	WMD
PPL	1	0.564	0.349	0.068	-0.406	-0.573	0.442
Gram. Errors	0.564	1	-0.495	-0.016	-0.413	**-0.629**	0.277
BLEU	0.349	-0.495	1	0.009	**0.796**	**0.798**	**-0.662**
Self-BLEU	0.068	-0.016	0.009	1	-0.022	-0.009	0.047
ROUGE	-0.406	-0.413	**0.796**	-0.022	1	**0.805**	**-0.767**
BERT	-0.573	**-0.629**	**0.798**	-0.009	**0.805**	1	**-0.700**
WMD	0.442	0.277	**-0.662**	0.047	**-0.767**	**-0.700**	1

The correlation analysis presented in table 2 uncovers significant relationships between the automatic evaluation metrics, enhancing our understanding of how they evaluate different aspects of text quality. The strong correlation among BERTScore, ROUGE-L, and BLEU indicates that these metrics focus on assessing text similarity. Specifically, there is a 0.796 correlation between BLEU and ROUGE-L, a 0.798 correlation between BLEU and BERTScore, and a 0.805 correlation between ROUGE-L and BERTScore. These metrics, which range from 0 to 1—where 0 indicates no similarity and 1 signifies an exact match—measure the similarity of the generated text to the reference text. Given their close correlations, it seems these metrics assess similar dimensions of text similarity. With correlation coefficients of -0.662, -0.767, and -0.700, respectively, Word Mover's Distance (WMD) shows a substantial negative connection with BLEU, ROUGE-L, and BERTScore. WMD is predicted to have a negative connection with similarity metrics since it measures the difference between generated and reference texts. Higher distance scores may result from generated texts that use diverse synonyms or paraphrases since WMD is especially sensitive to lexical differences. Although word order and lexical choices also have an impact on BLEU and ROUGE-L, the nature of their computations causes differences in their sensitivity. At -0.629, the number of grammatical errors and BERTScore are found to have a strong negative connection, meaning that larger grammatical errors are associated with lower BERTScore values. Furthermore, there is a moderately negative correlation between grammatical errors and ROUGE-L (-0.413) and BLEU (-0.495), indicating that texts with more grammatical errors typically score worse on these similarity metrics. It's interesting to note that there is a moderately positive association between grammatical errors and bewilderment (0.564), indicating that perplexity rises in tandem with grammatical errors. This is consistent with the idea that a language model has a tougher time predicting a sentence with poor structure. The fact that BERTScore and perplexity have a moderately negative association (-0.573) further supports the idea that lower-quality text with higher perplexity is probably going to have lower BERTScore values.

A number of metrics also correlate with perplexity, which gauges how effectively a language model predicts a specific word sequence. Lower prediction accuracy leads to more grammatical errors and a greater lexical distance from reference texts, as evidenced by its moderately positive link with BLEU (0.349), WMD (0.442), and grammatical errors (0.564). Perplexity, on the other hand, has a negative correlation with both BERTScore (-0.406) and ROUGE-L (-0.573), indicating that more resemblance to reference texts is a result of superior predictive performance. Perplexity and BLEU have a positive correlation, while ROUGE-L and BERTScore have a negative correlation. This discrepancy most likely results from the fact that BERTScore depends on contextual embeddings, ROUGE-L stresses lengthy common sequences, and BLEU concentrates on short n-gram overlaps. Consequently, a model that does well in text prediction might provide larger ROUGE-L and BERTScore values, but its BLEU score would not rise as sharply. There is no discernible relationship between any of the metrics and Self-BLEU, which gauges diversity in generated text. It has weak negative associations with grammatical errors (-0.016), ROUGE-L (-0.022), and BERTScore (-0.009), and weak positive correlations with perplexity (0.068), BLEU (0.009), and WMD (0.047). Self-BLEU's weak relationship to other measures suggests that text creation diversity is unrelated to predictive performance, lexical similarity, or grammatical precision. In conclusion, WMD shows a large negative connection with these metrics because of its inverse nature, whereas BLEU, ROUGE-L, and BERTScore all show strong correlations, demonstrating their efficacy in evaluating text similarity. Higher grammatical faults result in more bewilderment and poorer text similarity ratings. There are conflicting relationships between perplexity and text similarity, grammatical correctness, and prediction accuracy. Self-BLEU, meanwhile, maintains its independence, emphasizing its function in assessing diversity as opposed to fluency or resemblance.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Conclusion

Five language models are evaluated in this study using seven automated evaluation indicators, each of which highlights unique advantages and disadvantages. The results suggest that Word Mover's Distance, which gauges text differences, exhibits a negative connection with BLEU, BERT, and ROUGE scores, which all assess text similarity. There is a connection between perplexity and grammatical errors, with higher perplexity typically correlating to more grammatical issues. Furthermore, Self-BLEU, which measures the diversity of the generated text, does not show any correlation with the other metrics. The selection of evaluation metrics depends on the specific goal BLEU and ROUGE are effective for assessing text similarity, BERTScore and Word Mover's Distance focus on word diversity, and Self-BLEU is useful for measuring text diversity. Using a combination of several metrics leads to a more comprehensive evaluation, with human assessments further strengthening the reliability of results. The quality of text creation from the N-gram, CBOW, GRU, Pretrained GPT-2, and Finetuned GPT-2 models—which represent various phases of development—is compared for the first time in this work. This study provides in-depth explanations and pseudo-code for each metric, analyze their relationships and recommend an optimal strategy for metric selection. Future research will include human evaluations for evaluating coherence and redundancy, explore methods for sample selection to reduce the evaluation workload and compare human ratings with automated metrics to recognize meaningful correlations. Moreover, large-scale experiments on standard datasets will be carried out to refine text generation evaluation methods.

References

[1] J. Kim, Y. Heo, H. Yu, and J. Nang, "A Multi-Modal Story Generation Framework with AI-Driven Storyline Guidance," Electronics, vol. 12, no. 6, p. 1289, Jan. 2023, doi: https://doi.org/10.3390/electronics12061289.
[2] L. P. Khan, V. Gupta, S. Bedi and A. Singhal, "StoryGenAI: An Automatic Genre-Keyword Based Story Generation," 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2023, pp. 955-960, doi: 10.1109/CISES58720.2023.10183482.
[3] J. Valls-Vargas, J. Zhu, and S. Ontañón, "Towards End-to-End Natural Language Story Generation Systems," Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, vol. 13, no. 2, pp. 252–258, Jun. 2021, https://doi.org/10.1609/aiide.v13i2.12983.
[4] J.-W. Lin, Y.-C. Gao, and R.-G. Chang, "Chinese Story Generation with FastText Transformer Network," Feb. 2019, doi: https://doi.org/10.1109/icaiic.2019.8669087.
[5] A. Alabdulkarim, S. Li, and X. Peng, "Automatic Story Generation: Challenges and Attempts," arXiv (Cornell University), Jan. 2021, doi: https://doi.org/10.18653/v1/2021.nuse-1.8.
[6] Y. Wang, J. Lin, Z. Yu, W. Hu, and B. F. Karlsson, "Open-world story generation with structured knowledge enhancement: A comprehensive survey," Neurocomputing, vol. 559, pp. 126792–126792, Nov. 2023, doi: https://doi.org/10.1016/j.neucom.2023.126792.
[7] S.-M. Park and Y.-G. Kim, "Survey and challenges of story generation models - A multimodal perspective with five steps: Data embedding, topic modeling, storyline generation, draft story generation, and story evaluation," Information Fusion, vol. 67, pp. 41–63, Mar. 2021, doi: https://doi.org/10.1016/j.inffus.2020.10.009.
[8] A. I. Alhussain and A. M. Azmi, "Automatic Story Generation," ACM Computing Surveys, vol. 54, no. 5, pp. 1–38, Jun. 2022, doi: https://doi.org/10.1145/3453156.
[9] Sonali Fotedar, Koen Vannisselroij, S. Khalil, and B. Ploeg, "Storytelling AI: A Generative Approach to Story Narration.," International Joint Conference on Artificial Intelligence, pp. 19–22, Jan. 2020.
[10] J. Guan, F. Huang, Z. Zhao, X. Zhu, and M. Huang, "A knowledge-enhanced pretraining model for commonsense story generation", Trans. Assoc. Comput. Linguist., vol. 8, pp. 93–108, 2020.
[11] S. Goldfarb-Tarrant, T. Chakrabarty, R. Weischedel, and N. Peng, "Content planning for neural story generation with Aristotelian rescoring", in Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP), Online, Nov. 16–20, 2020, pp. 4319–4338.
[12] E. Clark and N. A. Smith, "Choose your own adventure: Paired suggestions in collaborative writing for evaluating story generation models", in Proc. 2021 Conf. North American Chapter Assoc. Comput. Linguist.: Human Language Technol. (NAACL), Online, Jun. 6–11, 2021, pp. 3566–3575.
[13] J. Guan, X. Mao, C. Fan, Z. Liu, W. Ding, and M. Huang, "Long text generation by modeling sentence-level and discourse-level coherence", in Proc. 59th Annu. Meeting Assoc. Comput. Linguist. and 11th Int. Joint Conf. Natural Language Process. (ACL-IJCNLP), Online, Aug. 2–5, 2021, pp. 6379–6393.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[14] L. Martin, P. Ammanabrolu, X. Wang, W. Hancock, S. Singh, B. Harrison, and M. Riedl, “Event representations for automated story generation with deep neural nets”, in Proc. AAAI Conf. Artif. Intell., New Orleans, LA, USA, Feb. 2–7, 2018.
[15] A. Fan, M. Lewis, and Y. Dauphin, “Strategies for structuring story generation”, in Proc. 57th Annu. Meeting Assoc. Comput. Linguist. (ACL), Florence, Italy, Jul. 28–Aug. 2, 2019, pp. 2650–2660.
[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., “Learning transferable visual models from natural language supervision”, in Proc. Int. Conf. Mach. Learn. (ICML), Online, Jul. 18–24, 2021, pp. 8748–8763.
[17] P. Li, Q. Huang, Z. Li, Y. Cai, F. Shuang and Q. Li, "Multi-Granularity Feature Fusion for Image-Guided Story Ending Generation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 3437-3449, 2024, doi: 10.1109/TASLP.2024.3419438.
[18] G. Wu, S. Guo, J. Hoffswell, G. Y. -Y. Chan, R. A. Rossi and E. Koh, "Socrates: Data Story Generation via Adaptive Machine-Guided Elicitation of User Feedback," in IEEE Transactions on Visualization and Computer Graphics, vol. 30, no. 1, pp. 131-141, Jan. 2024, doi: 10.1109/TVCG.2023.3327363.
[19] T. Rahman, H. -. Y. Lee, J. Ren, S. Tulyakov, S. Mahajan and L. Sigal, "Make-A-Story: Visual Memory Conditioned Consistent Story Generation," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 2493-2502, doi: 10.1109/CVPR52729.2023.00246
[20] Y. Xie, Y. Hu, L. Xing, Y. Li, W. Peng and P. Guo, "CLseg: Contrastive Learning of Story Ending Generation," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 8057-8061, doi: 10.1109/ICASSP43922.2022.9747435.
[21] A. Raza Samar, B. Khan and A. Mumtaz, "Context-Based Narrative Generation Transformer (NGen-Transformer)," 2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 2022, pp. 256-261, doi: 10.1109/IBCAST54850.2022.9990496.
[22] G. Chen, Y. Liu, H. Luan, M. Zhang, Q. Liu and M. Sun, "Learning to Generate Explainable Plots for Neural Story Generation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 585-593, 2021, doi: 10.1109/TASLP.2020.3039606.
[23] D. Shi, X. Xu, F. Sun, Y. Shi and N. Cao, "Calliope: Automatic Visual Data Story Generation from a Spreadsheet," in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, pp. 453-463, Feb. 2021, doi: 10.1109/TVCG.2020.3030403.
[24] J. -W. Lin, J. -H. Tseng and R. -G. Chang, "Chinese Story Generation Using Conditional Generative Adversarial Network," 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2020, pp. 457-462, doi: 10.1109/ICAIIC48513.2020.9065225.
[25] L. Wang, S. Qin, M. Xu, R. Zhang, L. Qi and W. Zhang, "From Quick-draw To Story: A Story Generation System for Kids’ Robot," 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO), Dali, China, 2019, pp. 1941-1946, doi: 10.1109/ROBIO49542.2019.8961449