Automated Narrative Craft: Exploring Machine Learning in Story Generation

Lakshmi Padmaja Dhyaram[0000−0001−8262−6821],
Anurag University, Hyderabad, India,
lakshmipadmajait@anurag.edu.in
Mrunalini Kota
Accenture, Hyderabad, India,
mrinalinikota001@gmail.com
Krishna Sriharsha Gundu,
Arizona State University, USA,
kgundu1@asu.edu
Surya Deepak G
Cognizant Technologies, Hyderabad, India.
deepak.surya.2001@gmail.com

ABSTRACT

With the advancements of Artificial Intelligence, we currently have NLP software like Grammarly that takes care of correctness, clarity, engagement of readers and also provides tone detection of text. However, they are only useful for grammatical purposes. For complex NLP tasks such as story generation, models such as machine translation, summarization or segmentation etc. has no relationship with the plot, events, actions and characters of a story. It purely focuses on the correctness and optimization of the given text. This led to the development of statistical and neural language models. Story generation by machines produces a story from a set of inputs using artificial intelligence. The aim of the paper is to examine numerous procedures used for Story Generation and to figure out the most competent procedure that can be adopted. This paper is a comparative study on some of the existing neural story generation algorithms. It seeks to identify the how far a story can be made interesting and relevant by neural networks.

Keywords - **Recurrent Neural Networks, Story Writing, Natural Language Processing, Machine Translation, Neural language models.**

Date of Submission: Date of Acceptance:

1. Introduction

Machines have effortlessly performed even the most difficult of computational tasks such as calculus, trigonometry, algebra, quantum physics etc. that humans find challenging. Recently, machine learning (ML) and natural language processing (NLP) technology has developed which made it feasible for machines to express creativity, originality and concision similar to humans. One such challenge that is studied and discussed in this paper is Story Generation [1].

A narrative encompasses a series of events, affective states, and psychological responses, which culminate in the transformative development of the characters in response to the unfolding storyline. The phenomenon under consideration comprises a sequence of occurrences that may be readily condensed into a concise summary. As individuals, people tend to conceptualize their thoughts and experiences in the form of narratives rather than just plotlines. These narratives encompass the various encounters and events that arise from our personal expectations and emotional states. A narrative encompasses the entirety of events experienced by the characters inside a certain fictional universe. The storyline encompasses a concise depiction of the events that are essential for the narrative to progress. The task of crafting a narrative can be challenging. Despite the abundance of social media platforms and networks designed to assist writers in their storytelling endeavors, authors frequently have difficulties in selecting appropriate language, experiencing a dearth of ideas, or struggling to initiate the writing process. These challenges, sometimes referred to as writer's block, can manifest in both fiction and non-fiction writing contexts[21].

Automated story generation is the use of artificial intelligence to produce a fictional story from a minimal set of inputs [9]. It finds many applications such as news generation from given phrases, assistant to help professional writers escape from writer's block, generating missing parts [5] of an incomplete story, story generation based on themes etc. Deep learning approaches proposed to solve this NLG problem rapidly increased over the last decade. Some of them are neural story generation algorithms. Many neural network based story generation techniques are coming up recently[8].

The first section of this paper gives a brief introduction on Story Generation. In section two, we discuss the existing types of neural story generation approaches. In section three, we compare different neural language models. Section four discusses controllable neural story generation. In the fifth section, we talk about other recent approaches to automated story generation[10][11]. The last section concludes our comparative study.

2. Related Work

Active research has been going on in the field of machine translation over the last couple of years, and also in the field of Automatic story generation. Statistical Machine Translation learns a probabilistic model from data. However, with the arrival of neural networks, Neural Machine Translation has also been explored. Then came the sequence-to-sequence (seq2seq) models, which involve two Recurrent Neural Networks (RNN)[24]. Its improved version is the use of attention mechanism. One of the drawbacks of neural language models is that the hidden state of the neural network (be it RNN or a transformer) only gives word choices based on the previous word tokens. Then came the concept of controllable story generation that focuses on plot points that have to be present in a story, in order to direct a story. When these proved to have their own limitations, algorithms that generate mental models of the readers were introduced[7]. This helped significantly improve story coherence.

A brief summary of the research papers that were examined for this study is given below. Table 1. shows the different methods applied for Automated Story Generation[12]. We mainly focused our study on story generation using neural networks. After going through these papers, we observed that there are various ways of approaching this NLG problem. No one method can be regarded as the best.

Table 1. Summarization of various methods applied for Automated Story Generation

S. No.	Techniques Used	Description
1.	Pb-SMT, syntax-SMT,seq2seq with attention-BiGRU encoder with 2-layer GRU decoder	Parag Jain et al., 2017, explored Statistical Machine Translation and Neural Machine Translation in automated story writing. They proposed a sequence-to-sequence recurrent neural network (seq2seq) architecture that used BiGRU as encoder and 2-layer GRU with attention mechanism as decoder. It takes in a set of short, textual descriptions as input and generates rich, humanlike stories as output. They also compared it with the existing SMT approaches, i.e., phrase based SMT and syntax-based SMT [16].
2.	Seq2Seq with attention-BiLSTM encoder with single LSTM decoder, RAKE algorithm, content-introducing method[3]	Lili Yao et al., 2019, proposed a plan-and-write framework that approaches the story generation into two problems -Story planning and Surface realization. To solve this two-step problem, they proposed a dynamic method based on content-introducing method. In this static schema based on seq2seq architecture, the title is encoded into a vector using a BiLSTM and words are generated in the storyline using another single-directional LSTM. They used RAKE algorithm to weigh the importance of words. Apart from existing metrics, they also proposed subjective metrics based on human evaluation of stories.
3.	Missing position prediction (MPP), Story completion (SC), Hierarchical-Seq2seq	Yusuke Mori et al., 2020, proposed a controllable neural story generation algorithm that deals with two problems, missing position prediction and story completion. Their proposed method consists of three parts- sentence encoder that uses SBERT algorithm, context encoder that uses GRU[20], and a BERT language model[18]. However, their study is only limited to 5 sentence stories[13].
4.	The Wordcraft editor- Evolved Transformer architecture.	Andy Coenen et al., 2021, introduced Wordcraft, an AI-assisted editor for story writing that uses Meena[22] dialog system to interact with the author and together come up with the plot of the story, get words down and rewrite existing text. It is compared with a traditional GPLM.
5.	Recurrent Neural Network Language Model (RNN LM)	Melissa Roemmele et al., 2018, demonstrated an interface that used RNN language model. It provides automated support for story writing, and suggestions for new sentences in a story. Their metrics were subjective, which was based on whether the author considered the suggestion given by the model or not. They adjusted the degree of randomness in the

		probability distribution and found out that as the generated suggestions varied, author's interactions were influenced[14].
6.	Transformer Neural Network(TNNs)	Kemal Araz. 2020, proposed a transformer network to improve quality of stories and help in correct pairing of prompts. It is then compared with the existing Hierarchical Neural Story Generation model[19]. Results showed that the transformer network could not outperform the state-of-art model. Even though it generated practical stories, it didn't pay much attention to the prompts of the generated stories [15].
7.	Story generation with Reader Models (StoRM)	In their study, Peng et al. (2021) introduced StoRM, a novel approach that use a commonsense inference model to establish correspondence between inferences formed at time t – 1 and those generated at time t. The system under consideration has four distinct modules, specifically: Knowledge Graph Building, Inference Exploration, Continuation Candidate Generation, and Graph Difference. The findings of the study indicated that StoRM generated narratives that exhibited a higher degree of coherence and goal-directedness compared to other methods.
8.	Explanatory Drama Generation And Recall (EDGAR) system- Para-COMET, ELI5 QA model, GPT-2.	Louis Castricato et al., 2021, proposed EDGAR using the question-answer model. It is an automated story generation system that constructs a story backwards from the ending sentence of a story. The system contains three major components: Question Generator that uses Para-COMET model, ELI5 QA model that generates answers to the questions describing one or more events, and a GPT-2 model that ranks the answers and chooses the best storyline which is added to the story.
9.	Seq2Seq, DRL-clustered, and DRL-unrestricted(Deep Reinforcement Learning)	Pradyumna Tambwekar et al., 2019 proposed a model based on Reinforcement Learning that extracts verb patterns from the story corpus, and clusters them, and then rewards the language model when it produces a continuation closer to the goal[23].
10.	CAST, COMET	Peng et al. 2021 proposed Commonsense-inference Augmented neural Story Telling (CAST), a framework that uses the COMET commonsense inference model to understand and analyze the characters in the story and match the wants and needs of the characters from previous events. Thus, it helps in achieving story coherence.

3. Neural Language Models

A language model calculates the probability of a token or sequence of tokens based on previously occurring tokens. It is trained on variety of story corpuses, from where it learns various ways of continuing a story. When the language model is sampled, it generates text related to the story [25]. When a prompt is given as an input, the language model generates one or more tokens as suggestions for the continuation of the story, which is again input into the model to progress the story further. That's how the model learns. Thus, sampling from a language model that is trained on large number of stories tends to produce text that looks like a story. Recently, Neural Network based language models were proposed to generate automatic stories. Some examples are Recurrent Neural Networks and Transformer Neural Network. BERT and other variants have also been applied in story generation. However, these language models only generate new tokens from previous ones. They cannot guarantee story coherence. Also, the cost of developing and maintaining these models is very high. Table 2. below compares the various existing neural language models.

Table 2. Comparison between various neural language models

Method	Advantages	Disadvantages
Statistical Machine Translation (SMT)	SMT uses Bayes Theorem to generate a probabilistic model where each sentence in the input is assigned with a probability. It translates words and phrases independently within a sentence. SMT systems produce higher BLEU/ METEOR/ROUGE values with pb-SMT being the best.	The current system lacks the necessary sophistication to construct a story that is semantically coherent with the input descriptions. The metrics used to evaluate them i.e., BLEU/ METEOR/ROUGE ignore the quality of text generated. So, these models are deviate from the story and the actual plot, however, they can provide a first-level insight. SMT is costly and laborious to develop because it needs lots of feature engineering and maintenance[3]
Sequence-to-sequence(Seq2Seq)	The employed methodology involves the utilisation of a singular neural network design known as sequence-to-sequence (seq2seq). This architecture has two recurrent neural networks (RNNs): an encoder RNN responsible for summarising the input, and a decoder RNN responsible for generating the target words. The performance of this alternative method surpasses that of SMT, while also necessitating significantly reduced maintenance.	It is less understandable, hard to debug and difficult to maintain. If the story training corpus is too diverse, producing unique sequences of sentences would be difficult.
Sequence-to-sequence(Seq2Seq) with attention	The attention mechanism significantly improves seq2seq. It provides more interpretability, and it's easier to debug by inspecting the attention distribution. While RNNs have long-term dependencies, with attention echanism, these are removed. That means each word in the target sequence only needs to find its match with few words in the source sequence, instead of matching with all the words of present in the sequence.	There are still many difficulties remaining, such as generated words being out-of-vocabulary or out of context, differences of genre in train and test data, and maintaining context as the story progresses.[3]
Transformer Neural Network(TNNs)	While sequential computation was used by seq2seq RNN models, TNNs provide pairwise parallelized multiplicative interaction (self-attention). This needed only one computation. It provides high performance with interpretability thereby producing state-of-the-art results. One of the advantages of Transformers is their capability to learn without the need for labeled data.	They suffer from plot incoherence as the story progresses and lack bsic commonsense and reasoning of humans. For instance, they focus on a single character and lose track of other characters. Transformers are expensive to develop and maintain, so it remains a privilege of the big technology companies with access to vast data sources and compute resources.
Recurrent Neural Networks (RNN LM)	When sentences are replaced with the built-in classes from WordNet or when verb frame classes are used from VerbNet, it theoretically reduces the sparsity RNN and LSTM and derivatives are able to learn a lot of long term information. They proposed a second neural network to convert event tuples into full sentences again to help progress the story further.	Initial models had trouble producing sentences that followed the story order due to sparsity[6]. It consumes a lot of time and resources to train these networks. They're inefficient and not scalable. They can remember sequences of 100s but not more.

4. Controllable Story Generation

Neural language models generate tokens based on the history of sequence of previous tokens, so they may generate incoherent stories that do not have a goal. These models cannot remember large sequences, so the earlier context may be forgotten. Even with larger context windows, the problem isn't solved altogether, because language models cannot build the story towards a future, except by accident.

But language models can also be tuned to focus on plot points that are necessary to be present in the story. This ensures the occurrence of specific events more probable as the story generation progresses. For instance, Angela fan et al., 2019 proposed a hierarchical fusion model that takes a one-sentence description of the story content and generates an entire paragraph. The plan and write technique[2] uses a two-level approach. It generates a sequence of keywords to plan a storyline. These are used to form and develop a language model which fabricates content about that keyword, thus generating stories based on each storyline. Experiments showed that their generated stories are more diverse, coherent, and on topic than the stories that were generated without creating a plan or a plot, according to both automatic and human evaluations. This schema is interactive and collaborative. Table 3. shows the framework of plan and write.

Table 3. Plan and write, Yao et al., 2019

Title (Given)	The Bicycle Accident
Storyline (Extracted)	Carrie → bicycle → sneak → nervous → leg
Story (Human Written)	Carrie recently acquired the skill of bicycle riding. The individual in question did not possess a bicycle that belonged to her. Carrie would surreptitiously avail herself of opportunities to ride her sister's bicycle. The individual had heightened anxiety while traversing an incline, resulting in a collision with a vertical structure. Carrie sustained a significant laceration on her leg as a result of the deformation of the bicycle frame.

The Missing Position Prediction(MPP) combined with Story Completion(SC) was proposed by Yusuke Mori , Hiroaki Yamane , Yusuke Mukuta , et al(2020). Given an incomplete story, it helps authors identify the location of the incomplete portion of a story. Their proposed system consists of three parts. Firstly, the sentence encoder receives the incomplete story as input and uses SBERT algorithm to generate sentence embeddings which are then sent to the context encoder that uses GRU[20] to generate a distributed representation of the entire context. For MPP, they input the context into a linear layer which generates a 5-unit output, since the model is trained on 5 sentence stories. For SC, the context is inputted into a BERT language model to complete the story. Their analysis shows that the accuracy of predictions when the missing part of a story is either at the beginning or the end is very high. However, their study is only limited to 5 sentence stories. They took it up as future work to use predicted MPP for SC. Fig 1. below demonstrates this model(from Fig 2).

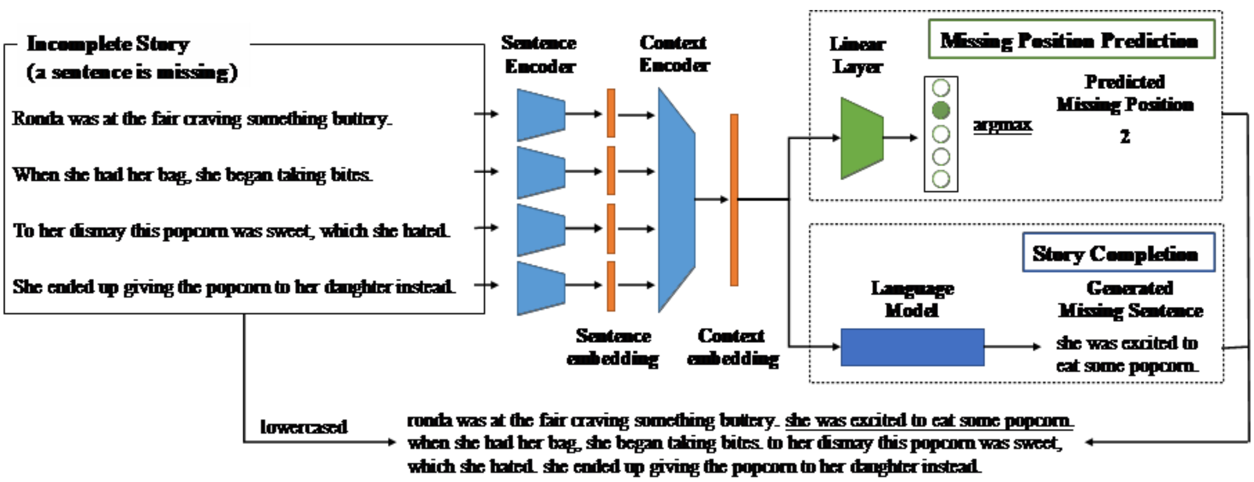

Fig 1. Missing Portion Prediction , proposed by Yusuke Mori et al.

Pradyumna Tambwekar et al., 2019 proposed a reward-shaping reinforcement learning technique, called DRL that examines a story corpus and generates intermediate rewards that are backpropagated into a pre-trained language model in order to guide the model in the direction of a given plotline [17]. They compared Seq2Seq, Deep Reinforcement Learning(DRL) restricted and unrestricted models. Upon experimentation, it was found that the Seq2Seq model mimicked the training corpus. Same was the case with the DRL-unrestricted model. However, the DRL-clustered model reached its goal quicker, while not jumping immediately to the goal.

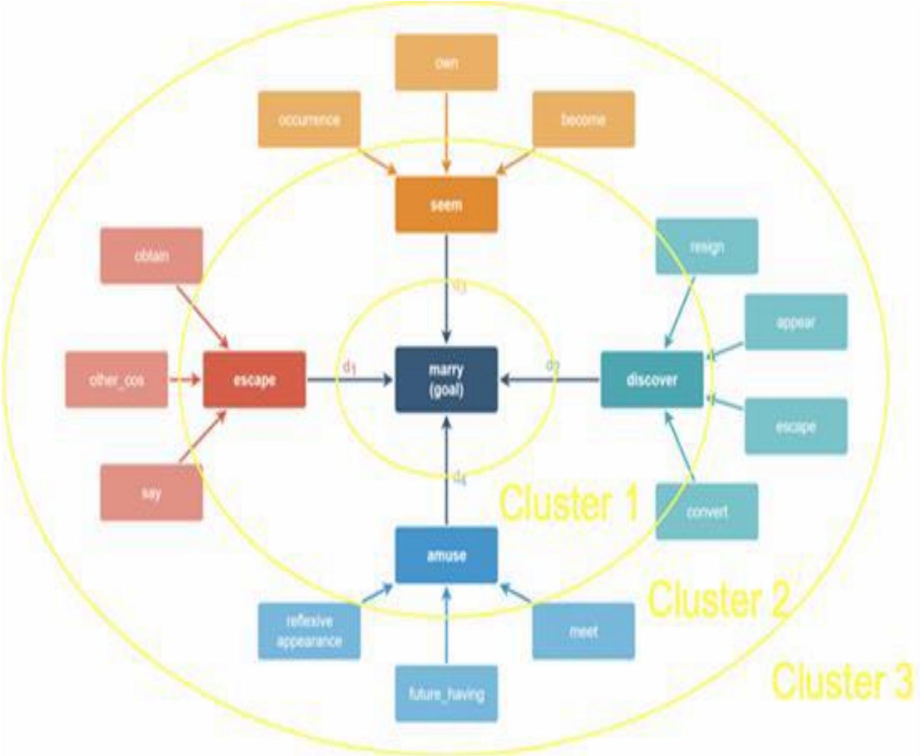

Fig 2. Demonstrates the algorithm.

5. Other Approaches

Another approach that has come up in the field of story generation is to pose queries to a language model that answers questions about the given story creation, where the answers to these carefully chosen questions become the content of the story itself. One such example is the Wordcraft text editor which has an integrated creative writing assistant[4]. Users start from a blank screen and as they write their story, they are provided with the option to take the help of the NLG assistant. The dialog system they used was Meena, a language model to help them come up with the sentences to continue the story and answer questions that posed in the form of a dialogue. Users can also try out their own queries by modifying the prompt. There are several interactions built into wordcraft, such as continuation of story, elaboration, rewriting the story in a different tone, etc. In minor qualitative studies, it was found that found Meena was just as effective as a similarly sized general-purpose language model (GPLM) that was not trained explicitly for dialog.

Another work is by Castricato et al., 2019, who proposed a system that generated a story backwards, starting with the end sentence of the story. The Para-COMET model in the system generates questions to deduce the series of events prior to the ending. They train an explanation generation language model(ELI5 QA) to answer the question and then send the generated answers to the GPT-2 model where answers are ranked and the best sequence added to the story. They then repeat this process. They inferred that this results in theoretically more coherent stories because each segment added to the story explains what comes next.

6. Conclusion

There came many changes in field of automated story generation. It developed from non-ML story generation systems to deep learning based story generation systems. Neural networks gave us the ability to acquire and make use of knowledge from large story datasets. Story generation systems that generate stories of various genre have been developed. But we should keep in mind about the psychology of readers to achieve story coherence. Even when the size of neural language models is increased, it cannot guarantee coherence in the stories generated. But we can't undermine the effectiveness of these approaches. The advantage of working on automated story generation is that there is no clear-cut path proven to be the best. New ideas are always welcomed.

References

1. Parag Jain, Priyanka Agrawal, Abhijit Mishra, Mohak Sukhwani, Anirban Laha, and Karthik Sankaranarayanan, Story generation from sequence of independent short descriptions, 2017.
2. Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan., Plan-and-write: Towards better automatic storytelling, 2019.
3. De-Yu, C. (Machine Translation - Towards Data Science. Medium.https://towardsdatascience.com/machine-translation-b0f0dbcef47c, January 6, 2022.
4. Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan., Wordcraft: a Human-AI Collaborative Editor for Story Writing, 2021.
5. Yusuke Mori, Hiroaki Yamane, Yusuke Mukuta, and Tatsuya Harada. ,Finding and generating a missing part for story completion 2021.
6. Melissa Roemmele and Andrew S. Gordon. Automated Assistance for Creative Writing with an RNN Language Model, 2018.
7. Xiangyu Peng, Kaige Xie, Amal Alabdulkarim, Harshith Kayam, Samihan Dani, Mark O. Riedl, Guiding Neural Story Generation with Reader Models, 2021.
8. Riedl, M., An Introduction to AI Story Generation - Mark Riedl. Medium. https://mark-riedl.medium.com/an-introduction-to-ai-story-generation-7f99a450f615, December26, 2021.
9. Kemal Araz., Transformer Neural Networks for Automated Story Generation. Master's Thesis. Technological University Dublin, Ireland, 2020.
10. Louis Castricato, Spencer Frazier, Jonthan Balloch, Nitya Tarakad, and Mark O. Riedl, Automated Story Generation as Question-Answering, 2021.
11. Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O Riedl., Automated storytelling via causal, commonsense plot ordering, 2020.
12. Amal Alabdulkarim, Siyan Li, and Xiangyu Peng. 2021. Automatic story generation: Challenges and attempts.
13. Culurciello, E., Notes on Transformer neural network for sequence to sequence learning. Medium. https://culurciello.medium.com/transformer-a-primer-da3f738441f, . April 16, 2019.
14. Culurciello, E., The fall of RNN / LSTM - Towards Data Science. Medium. https://towardsdatascience.com/the-fall-of-rnn-lstm-2d1594c74ce0, January 10, 2019a.
15. Angela Fan, Mike Lewis, and Yann Dauphin, Strategies for structuring story generation, 2019.
16. Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J Martin, Animesh Mehta, Brent Harrison, and Mark O Riedl., Controllable neural story plot generation via reward shaping, 2019.
17. Peng, X.; Li, S.; Wiegreffe, S.; and Riedl, M. O., Inferring the Reader: Guiding Automated Story Generation with Commonsense Reasoning, 2021.
18. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova., BERT: Pre-training of deep bidirectional transformers for language understanding, 2019.
19. Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical Neural Story Generation, 2018.
20. Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, Learning phrase representations using rnn encoder–decoder for statistical machine translation, 2014.
21. Kenitz, D., Writing Basics: Plot vs. Story. Skillshare Blog. https://www.skillshare.com/blog/writing-basics-plot-vs-story/ January 6, 2022.
22. Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le., Towards a Human-like OpenDomain Chatbot, 2020
23. Dickson, B., Artificial intelligence: Does another huge language model prove anything? TechTalks. https://bdtechtalks.com/2020/02/03/google-meena-chatbot-ai-language-model/, February 2 2020.
24. Brownlee, J., Gentle Introduction to Statistical Language Modeling and Neural Language Models. Machine Learning Mastery. https://machinelearningmastery.com/statistical-language-modeling-and-neural-language-models/, August 7, 2019.
25. Gundu, K.S., Dhyaram, L.P., Ramana Rao, G., Surya Deepak, G., Comparative Analysis of Energy Consumption in Text Processing Models. In: Rajagopal, S., Faruki, P., Popat, K. (eds) Advancements in Smart Computing and Information Security, Communications in Computer and Information Science, vol 1759. Springer, Cham. https://doi.org/10.1007/978-3-031-23092-9_9, ASCIS 2022.