ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

A Survey on Hierarchical Neural Story Generation

Sandeep Nithyanandan[1], Raseek C[2]

PG Student, Dept of Computer Science and Engineering, Government Engineering College, Palakkad, Kerala, India[1]

Assistant Professor, Department of Computer Science and Engineering, Government Engineering College, Palakkad, Kerala, India[2]

Abstract: Generating coherent and fluent text passages even using neural networks is still difficult. A story generation system using recurrent neural networks can be a solution to this. A large dataset of 300k human written stories along with writing prompts is used for hierarchical story generation. The system first generates a prompt and transforms it into a passage which ideally captures what the prompt intended to. The prompt is generated using a convolutional language model. The prompt gives a general sketch to what the story should be. A seq2seq model is used to generate a story that follows the prompt. A novel form of model fusion that improves the relevance of the story to the prompt, and adding a new gated multi-scale self-attention mechanism to model long-range context improves the coherence and the structure of the text passage. Also conditioning on the prompt makes it easier for the story to remain consistent and also have structure at a level beyond single phrase. A gated convolutional language (GCNN) is used for language model. A seq2seq model using LSTM and attention based decoder is used. Pre-training techniques have been used to improve the accuracy. Generating coherent and fluent passages has been improved efficiently using these techniques.

Keywords: Seq2seq Model, Hierarchical Story Generation, Gated Convolutional Language, Gated multi-scale self-attention mechanism

I. INTRODUCTION

Story-telling is on the frontier of current text generation technology: stories must remain thematically consistent across the complete document, requiring modelling very long range dependencies; stories require creativity; and stories need a high level plot, necessitating planning ahead rather than word by word generation [1]. Recent advances in machine learning based approaches for natural language generation have led to exploration of many diverse but related text generation tasks [6]. However, the existing systems/ approaches can be classified as weak AI systems. According to the classical definition, a strong AI based nlg system should perform language generation in the same manner, expressing similar levels of creativity, originality and brevity as humans.

The method tackles the challenges of story-telling with a hierarchical model, which first generates a sentence called the prompt describing the topic for the story, and then conditions on this prompt when generating the story. Conditioning on the prompt or premise makes it easier to generate consistent stories because they provide grounding for the overall plot. It also reduces the tendency of standard sequence models to drift off topic. To improve the relevance of the generated story to its prompt, a fusion mechanism is introduced where the model is trained on top of a pre-trained seq2seq model. To improve over the pre-trained model, the second model must focus on the link between the prompt and the story. The fusion mechanism show that it can help seq2seq models build dependencies between their input and output.

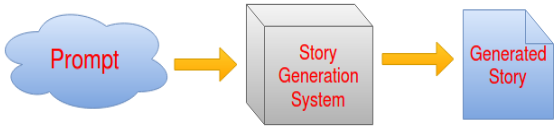


Fig. 1. Story Generation System

The main objective of this survey is to identify the different techniques that are used for the process of Natural Language Generation , mainly Story Generation. The process of identifying the best technique that generates the story which adheres to the central plot of the story is essential. The identification of dataset which aids in the story generation

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

is another objective of the work. The work identifies the technique which generates a story adhering to the central plot and a proper dataset for the generation process.

The problem of story generation introduces a variety of challenges which needs to be tackled [1].

- Standard sequence-to-sequence (seq2seq) models applied to hierarchical story generation are prone to degenerating into language models that pay little attention to the writing prompt. This failure is due to the complex and under specified dependencies between the prompt and the story, which are much harder to model than the closer dependencies required for language modelling.
- Another major challenge in story generation is the inefficiency of modelling long documents with standard recurrent architectures stories contains 734 words on average in the dataset. To improve efficiency a convolutional architecture is used, allowing whole stories to be encoded in parallel. Existing convolutional architectures only encode a bounded amount of context.

Automatic story generation has a long-standing tradition in the field of Artificial Intelligence. The ability to create stories on demand holds great potential for entertainment and education. The story generated by the machines shows the creativity the machines have achieved. Automatic Story Generation explores the creative side of the machines. There are various levels of application for it. Some important fields are:

- Computer Game Environment - Modern computer games are becoming more immersive, containing multiple story lines and hundreds of characters. This has substantially increased the amount of work required to produce each game. However, by allowing the game to write its own story line, it can remain engaging to the player [7].
- Education - Intelligent tutoring systems can potentially provide students with instant feedback and suggestions of how to write their own stories.
- Soap Operas - Creative system that can generate story based on situation can be effectively used to create stories for different Soap Operas.

In this paper it describes different approaches for the task of story generation. This paper is organized as follows: Section II gives a formal definition of story generation. Section II also discusses about various classification (machine learning) approaches, the same and provides a comparison between them. Section IV gives a brief concluding comments.

II. HIERARCHICAL STORY GENERATION

Nearly 20 years since its first commercial application, Natural Language Generation (NLG) has made a name for itself and established its own vertical across many industries and use cases. The ability to automatically turn data into clear, natural language has transformed the way companies and organizations interact with and act upon their data. The problem of story generation can be formally defined as:

Given a prompt p, the goal of the story generation is to generate a story for p using hierarchical methods.

A. Different Approaches on story Generation

There are different approaches for the task of story generation. In Ilya Sutskever. et.al[2], they present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure [2]. The method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. The Recurrent Neural Network (RNN) is a natural generalization of feed forward neural networks to sequences. Given a sequence of inputs (x1, ..., xT), a standard RNN computes a sequence of outputs (y1, ..., yT) by iterating the following equation:

The model is differing in three important ways:

- Use two different LSTMs: one for the input sequence and another for the output sequence, because doing so increases the number model parameters at negligible computational cost and makes it natural to train the LSTM on multiple language pairs simultaneously.
- Deep LSTMs significantly outperformed shallow LSTMs, so here chose an LSTM with four layers.
- Find extremely valuable to reverse the order of the words of the input sentence.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

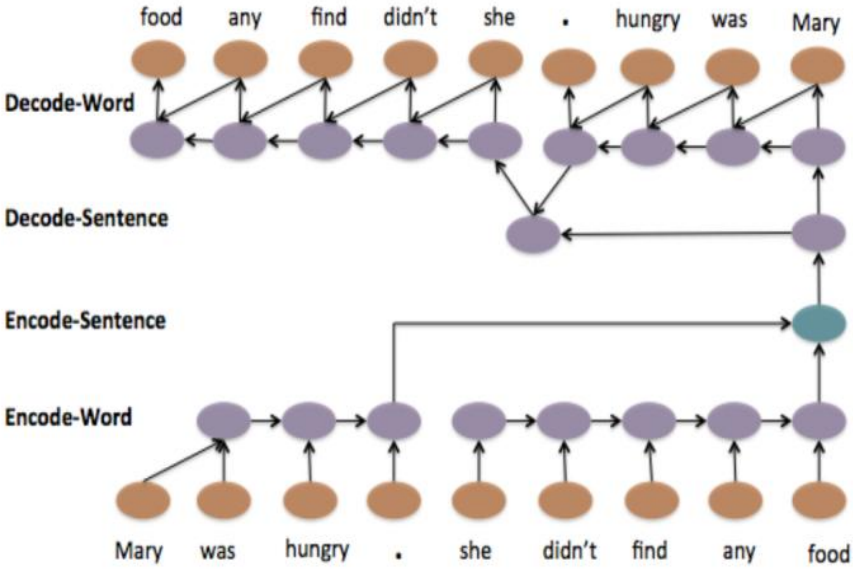


Fig. 2. Hierarchical Sequence to Sequence Model with Attention

In Ilya Sutskever et.al[2] the WMT14 English to French dataset is used. Train models on a subset of 12M sentences consisting of 348M French words and 304M English words, which is a clean selected subset from. Chose this translation task and this specific training set subset because of the public availability of a tokenized training and test set together with 1000-best lists from the baseline SMT.

The work proposed by Jiwei Li et.al. [3] explore an important step toward this generation task: training an LSTM (Long-short term memory) auto-encoder to preserve and reconstruct multi-sentence paragraphs. They introduce an LSTM model that hierarchically builds an embedding for a paragraph from embeddings for sentences and words, then decodes this embedding to reconstruct the original paragraph [3]. Model 1 deals with standard LSTM, in which whole input and output are treated as one sequence of tokens. Trained an auto-encoder that first maps input documents into vector representations from a LSTM encode and then reconstructs inputs by predicting to- kens within the document sequentially from a LSTM decode. Model 2 deals with Hierarchical LSTM, draws on the intuition that just as the juxtaposition of words creates a joint meaning of a sentence, the juxtaposition of sentences also creates a joint meaning of a paragraph or a document. Model 3 deals with Hierarchical LSTM with Attention, Attention models adopt a look-back strategy by linking the current decoding stage with input sentences in an attempt to consider which part of the input is most responsible for the current decoding state.

The same work by [3] Implemented the proposed auto-encoder on two datasets, a highly domain specific dataset consisting of hotel reviews and a general dataset extracted from Wikipedia. Here use a subset of hotel reviews crawled from Trip Advisor. Consider only reviews consisting sentences ranging from 50 to 250 words; the model has problems dealing with extremely long sentences. From Wikipedia dataset, extracted paragraphs from Wikipedia corpus that meet the aforementioned length requirements. Paragraphs with larger than 4 percent of unknown words are discarded.

In work by Ramachandran .et.al[4] presents a general unsupervised learning method to improve the accuracy of sequence to sequence (seq2seq) models [4]. The weights of the encoder and decoder of a seq2seq model are initialized with the pretrained weights of two language models and then fine-tuned with labelled data. This method applies to challenging benchmarks in machine translation and abstractive summarization and finds that it significantly improves the subsequent supervised models. In sequence to sequence learning, an RNN encoder is used to represent x1, ..., xm as a hidden vector, which is given to an RNN decoder to produce the output sequence. This method is based on the observation that without the encoder, the decoder essentially acts like a language model on ys. Similarly, the encoder with an additional output layer also acts like a language model. Thus it is natural to use trained languages models to initialize the encoder and decoder. Therefore, the basic procedure of this approach is to pretrain both the seq2seq encoder and decoder networks with language models, which can be trained on large amounts of unlabeled text data.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

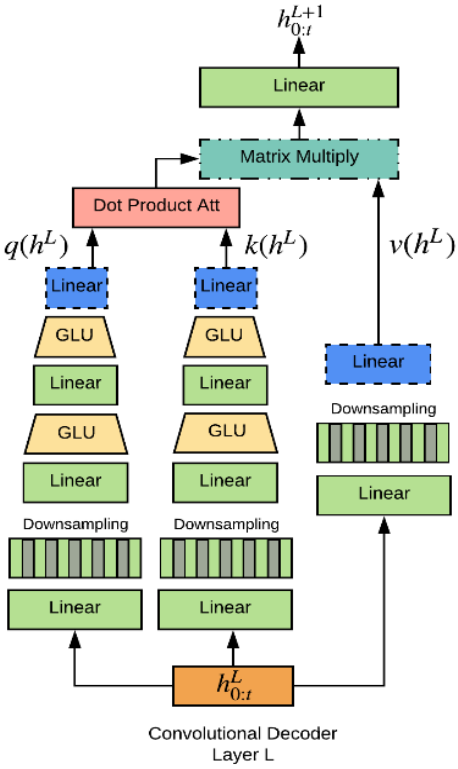
PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

The dataset in [4], For machine translation, evaluate the method on the WMT English German task. used the WMT 14 training dataset, which is slightly smaller than the WMT 15 dataset. Because the dataset has some noisy examples, they used a language detection system to filter the training examples. Sentences pairs where either the source was not English or the target was not German were thrown away. This resulted in around 4 million training examples.



Authors in [5] explore story generation: creative systems that can build coherent and fluent passages of text about a topic. The model first generates a premise, and then transforms it into a passage of text [1]. The work gain further improvements with a novel form of model fusion that improves the relevance of the story to the prompt, and adding a new gated multi-scale self-attention mechanism to model long-range context. High-level structure is integral to good stories, but language models generate on a strictly-word-byword basis and so cannot explicitly make high-level plans. First, generate the premise or prompt of the story using the convolutional language model. The prompt gives a sketch of the structure of the story. Second, they use a seq2seq model to generate a story that follows the premise. Conditioning on the prompt makes it easier for the story to remain consistent and also have structure at a level beyond single phrases. The two different approaches used are;

1) *Efficient Learning with Convolutional Sequence-to- Sequence Model:* To transform prompts into stories, instead of build on the convolutional seq2seq model, which uses deep convolutional networks as the encoder and decoder? Convolutional models are ideally suited to modelling long sequences, because they allow parallelism of computation within the sequence. In the Conv-seq2seq model, the encoder and decoder are connected with attention modules that perform a weighted sum of encoder outputs, using attention at each layer of the decoder.

2) Modelling Unbounded Context with Gated Multi-Scale Self-attention: CNNs can only model a bounded context window, preventing the modelling of long-range dependencies within the output story. To enable modelling of unbounded context, supplement the decoder with a self-attention mechanism, which allows the model to refer to any previously generated words. The self-attention mechanism improves the models ability to extract long-range context with limited computational impact due to parallelism.

3) *Gated Attention*: Use multi-head attention to allow each head to attend to information at different positions. However, the queries, keys and values are not given by linear projections but by more expressive gated deep neural nets with Gated Linear Unit activations. They show that gating lends the self-attention mechanism crucial capacity to make fine-grained selections.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

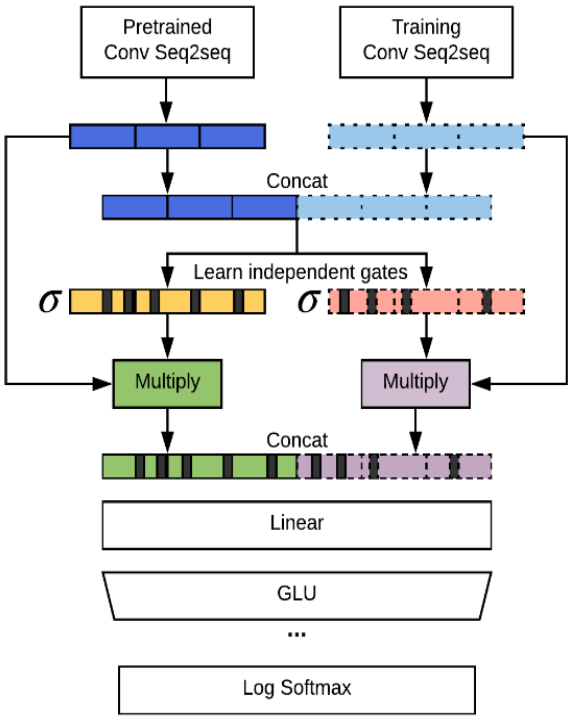
National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

4) *Multi-Scale Attention*: Each head operating at a different time scale, depicted in Figure 2.5. Thus the input to each head is down sampled a different amount the first head sees the full input, the second every other input time step, the third every third input time step, etc. The different scales encourage the heads to attend to different information. The down sampling operation limits the number of tokens in the attention maps, making them sharper.

5) *Improving Relevance to Input Prompt with Model Fusion* : Seq2seq models ignore the prompt and focus solely on modelling the stories, because the local dependencies required for language modelling are easier to model than the subtle dependencies between prompt and story. Here propose a fusion-based approach to encourage conditioning on the prompt. They train a seq2seq model that has access to the hidden states of a pre-trained seq2seq model. Doing so can be seen as a type of boosting or residual learning that allows the second model to focus on what the first model failed to learn such as conditioning on the prompt.



Angela Fan .et.al. [5] used a hierarchical story generation dataset from Reedits WRITING PROMPTS forum is used. WRITING PROMPTS is a community where online users inspire each other to write by submitting story premises, or prompts, and other users freely respond. Each prompt can have multiple story responses. The prompts have a large diversity of topic, length, and detail. The stories must be at least 30 words, avoid general profanity and inappropriate content, and should be inspired by the prompt.

CONCLUSION AND FUTURE WORK

This survey compares different methodologies for story generation. The major works for story generation uses Recurrent Neural Networks for sequential learning. The effectiveness of Convolutional Neural Networks(CNN) for story generation is now being explored. Hierarchical Learning methods are being effectively proposed to capture the dependencies within the sentences being generated. A combination of of Hierarchical seq2seq learning with CNN outperforms all the other method that performs story generation. The methods can also use a premise or ground plot for generating the story, also to avoid drifting off-topic while story generation. Another insight from the survey shows the importance of dataset for the process of story generation. A quality dataset effective story generation is a necessary need to be addressed. With the use of quality dataset and effective learning techniques like Hierarchical learning models the creativity of the machines can be improved, thereby making them to generate coherent, fluid stories.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

REFERENCES

[1]. Angela Fan, Mike Lewis, Yann Dauphin.Hierarchical Neural Story Generation, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018.
[2]. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks.In Neural Information Processing Systems(NIPS), 2017.
[3]. Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. arXiv 2017.
[4]. Prajit Ramachandran, Peter J Liu, and Quoc V Le. Unsupervised pretraining for sequence to sequence learning. arXiv 2017
[5]. Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. Cold fusion: Training seq2seq models together with language models. arXiv, 2017.
[6]. Denis Yarats and Mike Lewis.Hierarchical text generation and planning for strate- gic dialogue. arXiv, 2017.
[7]. Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modelling with gated convolutional networks, 2017.
[8]. Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei- Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. On using monolingual corpora in neural machine translation. arXiv, 2015.
[9]. Brent Harrison, Christopher Purdy, and Mark O Riedl. Toward automated story generation with markov chain monte carlo methods and deep neural networks, 2017. [
[10]. Xingxing Zhang and Mirella Lapata. Chinese poetry generation with recurrent neural networks. In Proceedings of the 2014 Conference on Em-pirical Methods in Natural Language Processing (EMNLP), 2014.
[11]. Melissa Roemmele. Writing stories with help from recurrent neural networks. In AAAI.2016.
[12]. Alexander M Rush, Sumit Chopra, and Jason We- ston. A neural attention model for abstractive sentence summarization. arXiv 2015.
[13]. Lara J Martin, Prithviraj Ammanabrolu, William Han- cock, Shruti Singh, Brent Harrison, and Mark O Riedl. Event representations for automated story generation with deep neural nets. arXiv2017.
[14]. Jiwei Li, Minh-Thang Luong, and Dan Juraf- sky. A hierarchical neural autoencoder for paragraphs and documents. arXiv 2015.