SI 630 WN24 Project Proposal
Short Story Ending Generation

Divya Santhanam and **Divyam Sharma**

1 Project Goals

Writing stories is an engaging yet challenging endeavor. Often, authors encounter moments of creative block, where the path forward in their narrative becomes obscured. This project is designed to address such moments by providing an innovative solution: a tool that completes stories based on given prompts. By inputting a short story prompt, users can receive a conclusion to their story, articulated in one sentence or more, thereby enhancing the storytelling process with AI-driven creativity. This tool aims not only to assist authors in navigating writer's block but also to offer a fun and interactive way for anyone to expand on story ideas spontaneously. Through this project, we explore the intersection of artificial intelligence and creative writing, pushing the boundaries of how stories can be crafted and concluded.

2 NLP Task Definition

As discussed in the Project Goals section, our project aims to develop a sophisticated text-generator that effectively writes an applicable and readable conclusion to a short story prompt.

Our text-generator will require four sentences of text input that represent the bulk of the story. From this text, our generator will produce a one-sentence-long conclusion. The generator will be trained on thousands of five-sentence-long stories that resemble the type of user input that we would expect to see.

Additionally, once our model has been trained and performs well as determined by several readability metrics, we plan to conduct additional training on 30-40 longer stories to see how our model performs on generating longer conclusions.

Our project aims to solve the problem of writer's block. There are several instances where having a tool that can conclude a story would be beneficial. For example, a potential user of our text-generator could be an author looking for a conclusion to their next short story. Another use case could be a parent that is looking for a personalized bedtime story to tell their child.

3 Data

The data we will be using for our project will be five-sentence-long short stories. To obtain the short stories, we will be using a combination of both preexisting online databases as well as minimal web scraping. As of the due date of the Project Proposal, we have already obtained the bulk of the data we will be using for the core focus of our training.

The primary dataset we plan to use is from the Story Cloze Test and ROCStories Corpora which has approx. 95,000 short five-sentence-long stories that can be found here. The dataset has free access to anyone upon request, which we have already received. This project aims to evaluate story prediction. Given two possible conclusions to a story, the model chooses the more favorable option. Our project extends this concept by focusing on text generation over prediction. Rather than being given options to choose from, our model will generate the text itself.

Additionally, we will be sourcing several stories from the American Literature short story collection that can be found here. These stories will be used after training our model to work with single-sentence generation. If we find that our data is not sufficient, we may also use the Short Story Index that can be found here which can be accessed by free trial.

4 Related Work

A similar work that has been done by (Mostafazadeh et al., 2016) at the University of Rochester has focused on coming up with a benchmark - Story Cloze benchmark

as a method for evaluating the understanding of short (five-sentence) commonsense stories. In the Story Cloze benchmark, the first four sentences of an everyday story are given, and the concluding sentence is withheld. The model is offered two choices, both of which are contextual to the preceding sentences but one of which is right and the other is wrong. So the focus of the above-mentioned work is on predicting the right ending out of the two options. Other works on similar lines focusing on predicting the right ending to solve ROCStory Cloze Task have been such as (Schwartz et al., 2017). To improve the performance, features like topic words and sentiment score are also extracted and incorporated (Chaturvedi et al., 2017). Neural network models have also been applied to this task (e.g., (Huang et al., 2013) and (Cai et al., 2017)), which use LSTM to encode different parts of the story and calculate their similarities. In addition, (Li et al., 2018) introduces event frame to their model and leverages five different embeddings. Finally, (Radford et al., 2018) develops a transformer model and achieves state-of-the-art performance on ROCStories, where the transformer was pre-trained on BooksCorpus (a large unlabeled corpus) and finetuned on ROCStories.

Our intended work will focus on being able to predict the whole last line of the 5-sentence story, aiming to have contextual meaning and logical coherence with the rest of the story (the 4 seen sentences) and not just selecting the right ending choice out of a given few options.

5 Evaluation

Evaluating story-ending generation requires a nuanced approach to gauge both the technical and creative quality of generated narratives. Therefore, we introduce here various evaluation methods which we plan to test our model on, encompassing automated metrics like BLEU, ROUGE, METEOR, BERTScore, and perplexity, which quantitatively measure aspects such as n-gram overlap, semantic similarity, and fluency. Additionally, human evaluation will be important for looking into coherence, creativity, and emotional impact that automated metrics might overlook. Together, these evaluation strategies will provide us with a comprehensive assessment framework, ensuring a balanced analysis of a model's ability to generate compelling and contextually fitting story endings.

5.1 BLEU Score (Papineni et al., 2002)

To apply the BLEU score for evaluating a story-ending generation task where we have the actual 5th sentence of the story and a generated 5th sentence, we'll compare the generated sentence against the actual sentence using the BLEU metric. BLEU will quantify how close our generated ending is to the actual ending based on n-gram overlap.

5.2 ROUGE (Lin, 2004)

The ROUGE Score and its variants (Rouge-N/L/W) facilitate evaluation by measuring the overlap of n-grams, the longest common subsequences, and skip-bigrams between the generated text and a set of reference texts. For story ending generation, ROUGE can assess the extent to which key phrases and narrative elements in the generated story ending align with those in the actual story ending, offering a quantitative measure of narrative fidelity.

5.3 METEOR (Banerjee and Lavie, 2005)

In the context of story ending generation, METEOR provides a nuanced assessment of how well the generated ending captures the meaning and fluency of the actual story ending, taking into account paraphrasing and flexible expression.

5.4 BERTScore (Zhang et al., 2019)

BERTScore leverages the contextual embeddings from models like BERT to compute the semantic similarity between the generated text and the reference text. By comparing the cosine similarity of embeddings for matched words, BERTScore offers an evaluation of semantic congruence between the generated story ending and the actual ending, highlighting the model's ability to generate semantically relevant and contextually appropriate narrative conclusions.

5.5 Perplexity (Jelinek et al., 1977)

Perplexity measures the uncertainty of a language model in predicting the next token, providing an indication of the fluency and naturalness of the generated text. For story ending generation, a lower perplexity score on the generated ending suggests that the text is more predictable and fluent, reflecting the model's linguistic competency in crafting coherent and contextually fitting narrative closures.

5.6 Human Evaluation

Despite the advancement in automated metrics, human evaluation remains indispensable for assessing

creative text generation tasks such as the story ending generation. Human judges - the authors of this work - Divya and Divyam will each evaluate 500+ one sentence generated endings for coherence, narrative satisfaction, creativity, emotional impact, and grammatical correctness, looking deeply into the qualitative aspects of story generation that automated metrics cannot fully capture.

6 Work Plan

For the following two weeks after submitting our proposal, both members of our team will read through the papers mentioned in the Related Works section and look for commonalities that can help guide our decision making process. Additionally, we will both collect any remaining necessary data from the sources outlined in the Data section so that we will be prepared to start training our generator. In the next two weeks, we will create unigram and bigram models that are trained on five-sentence stories as a baseline to test the performance of our generator. We will evaluate our models using a variety of readability scores outlined in the Evaluation section. Once we have a working baseline, we will create more complex models after the Project Update.

As a two person group, we plan on creating a more complex project by using text generation rather than text prediction. Our evaluation system will therefore need to be more robust, and we will execute this by using a variety of readability metrics. Both members of our team will implement several of these scoring algorithms. As another challenge in our project, we plan on conducting an additional round of training using longer stories to evaluate our model's performance on text-generation for longer paragraphs. This training will involve about 30-40 additional stories.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending: Strong neural baselines for the roc story cloze task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 616–622.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story comprehension for predicting what happens next. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1603–1614.

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

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Qian Li, Ziwei Li, Jin-Mao Wei, Yanhui Gu, Adam Jatowt, and Zhenglu Yang. 2018. A multi-attention based neural network with external knowledge for story ending predicting task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1754–1762.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A Smith. 2017. Story cloze task: Uw nlp system. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 52–55.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.