

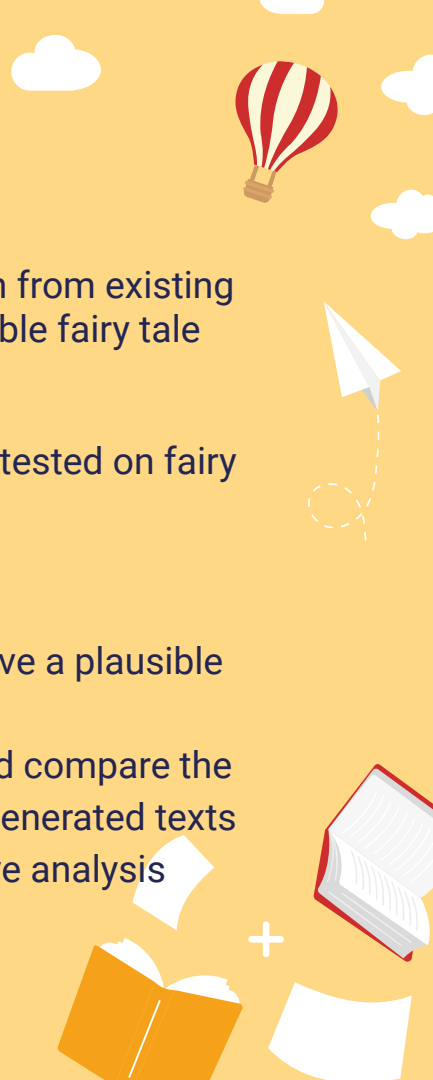
Project Breadcrumbs: Fairy Tale Text Generation

Andrew Arteaga, Maddie Louis, & Merissa Tan



GOALS

- Create a text generation model that will learn from existing fairy tales and auto-generate new and plausible fairy tale texts
- Quantitative analysis
 - Compare perplexities of models when tested on fairy tale texts
 - Aim to minimize perplexity
- Qualitative analysis
 - Aim for generated fairy tale texts to have a plausible story / plot line
 - Incorporate author-specific models and compare the similarities and differences of model-generated texts
 - Use word clouds to help with qualitative analysis



APPROACH

GPT-2 is already pre-trained!

Baseline model

- We fine-tune the model on fairy tale train data and compare the perplexity of the original model vs. the fine-tuned model when tested on fairy tale texts

Final model

- We create different language models, each trained for a specific author
 - Only fine-tuned on specific author dataset
 - Fine-tuned on a large fairy tale dataset **and** a specific author dataset
- For each model, compare perplexity of two different versions on that author's text



DATA AND MODELS

Baseline model

The Blue Fairy Book - Andrew Lang

Final model

Author-specific datasets

- Charles Perrault
- Grimms Brothers
- Hans Christian Andersen

Large dataset consisting of many different datasets, including "The Blue Fairy Book" and "Japanese Fairy Tales" by Yei Theodora Ozaki

Language /
Generation Models:
GPT-2

Data

Models



DEMO



BASELINE RESULTS

Test Dataset Used	Baseline Perplexity	Fine-tuned Perplexity	Fine-tuned Generation Samples
Blue Fairy Book	41.0371	29.3047	The Prince took her in his arms and kissed her gently; but when he came to the Fairy he thought he would have the very most pleasant thing to do. In spite of all the fuss he found his way to the Fairy's house. But there was nothing there except a long staircase, which was covered with flowers.
Hans Christian Andersen	62.7144	43.4158	And he said to him, "Look at the little boy! He looks like a lion; if you can look at it, your daughter will love him as she loves her brother." That was a very unhappy expression, for, indeed, every time one spoke to him, there was a change in his complexion; the tears of happiness came again in the little brother.
Kaggle Fairy Tale Dataset	60.0117	45.2276	But what they could not tell the king, and how his sister had betrayed them, the princesses and the princes felt compelled to tell the old lady that they had no idea what had happened, and what to do with the prince who had promised their daughter to a beautiful woman. It was not till the queen's daughter had taken them into a great palace that they felt free to marry them themselves.

- Results were as we expected:
 - Fine-tuned perplexity is lower than the baseline
 - Fine-tuned generated texts showed obvious usage of fairy tale language and motifs



ALLOCATION OF WORK



- Divide and Conquer Approach:
 - Each group member created their own baseline and fine-tuned model
 - Each member researched and parsed a large generic fairy tale dataset and a fairy tale dataset that followed a specific author style
 - Each member completed an equal amount of the written work
- Reasoning:
 - It was concluded to be a fair division of work because it gave everyone a chance to understand how the language/generation model worked without them having to train and fine tune each and every single model on a different data set






WHAT WE LEARNED SO FAR



- 
- The model that's fine-tuned on a general fairy tale dataset AND an author-specific dataset has a slightly lower perplexity than a model that's only fine-tuned on the latter
 - The model performs significantly better with Grimms' fairy tales compared to other authors
 - Fairy tales generated after fine-tuning on a specific author are more plausible than fairy tales generated on a conglomeration of fairy tales from different authors
- 



REMAINING GOALS

- Potentially add more fairy tale texts to our large combined dataset
 - Requires more manual work
 - Need to ensure the generic fairy tale dataset doesn't contain any stories from the specific authors we are using, to prevent overfitting
 - Minimize perplexity for all of our models
 - Find other ways to tune the models, such that they generate more plausible stories
 - ex. Would adding more data lead to more plausible stories?
- 

Questions?

Thanks for listening!

