Text Generation: Story Ending Prediction

Could T5 model understand the causal relationship and generate reasonable story endings?

Fengyao Luo
fengyaoluo@berkeley.edu

Ming Chen
mingchen@ischool.berkeley.edu

Abstract

Story telling by a machine has fascinated many science fiction writers. With the development of technology such as GPT-3, BERT, and T5, machines can generate reasonable and fluent sentences with certain guidance. However, it is still a challenge for machines to understand the causal relationship between events and understand the related ideas within sentences.

This research explored the generation function from the latest Encoder-Decoder Model T5, and applied 2 different sentence similarity methods (T5 sentence similarity, Universal Sentence Encoder) to evaluate the model performance on the Story Cloze Test. We achieved the baseline val accuracy as 71.4%. Error analysis revealed that story ending generation varied and similarity scores between output and ending 1 or ending 2 are very close. Furthermore, we trained the model to output 5 endings and applied a Simi-Senti score (sentiment consistency indicator * 1 + similarity score) to the model, which improved model performance by 6.6%, and reached the final validation accuracy of 76.1%. We reached a test accuracy of 74.5% on the leaderboard of Story Cloze Test Winter 2018.

1 Introduction

With artificial intelligence becoming increasingly popular, people are wondering whether machines can understand language and generate stories smoothly. In recent years, GPT-3[2], BERT[4], and T5[13] have been introduced to markets, demonstrating that they have the capability to tell stories with given topics. However, scholars argued that machines still have difficulties understanding the inner logics and coherence in corpses (Elazar, Yanai et al., 2019)[5] . Even though machines can generate language texts which look reasonable at the first glance, the cohesion and coherence in the language is still a challenge to the AI industry.

ROCStories Corpora[10], which is a new corpus of five-sentence commonsense stories (Appendix A). Story Cloze Test[16] is an ideal framework to understand the machine's capability of script reading, as it asks the machine to pick the right story ending, which follows the flow of the story (Appendix B).

Previous studies applied various models to solve the problem of improving machine logic, including BERT (Li, Ding, Liu, 2019)[7], RNN binary classification model (Roemmele, Kobayashi, et al., 2019)[14], and event based neural network (Martin, Lara J et al., 2017)[9]. This research introduced a new way to tackle the ROC Story Cloze Test, which fine-tuned T5 on ROCStory Corpora to generate the next sentence and then use the similarity score to vote for the correct answer between right and wrong endings in the Story Cloze Test.

This paper makes the following contributions: 1) We applied unsupervised learning and fine-tuned pretrained T5 model on over 50,000 stories, which can be useful for generating reasonable stories and generalizing the free text situation. 2) This research presents the Simi-Senti scores (sentiment consistency indicator * 1 + similarity score) during inference on the Story Cloze Test. 3) Based on the experiments, we found out that increasing the number of sequence outputs and applying sentiment score during inference is helpful in improving accuracy.

2 Related Work

Previous studies have basically taken two approaches: 1) Supervised learning approach, applied transfer BERT to fine-tune as a classifier, which was trained on the SCT v1.0 validation set, and then evaluated on the SCT v1.5 validation, and obtained an accuracy of 91.8% (Li, Ding, X., Liu, T. , 2019)[7]. 2) Unsupervised learning approach, the best performance has been conducted by a RNN-based model with the negative samples as augmentation (Roemmele, Kobayashi, et al., 2019)[14]. The augmentation was conducted by 4 ways: ran-

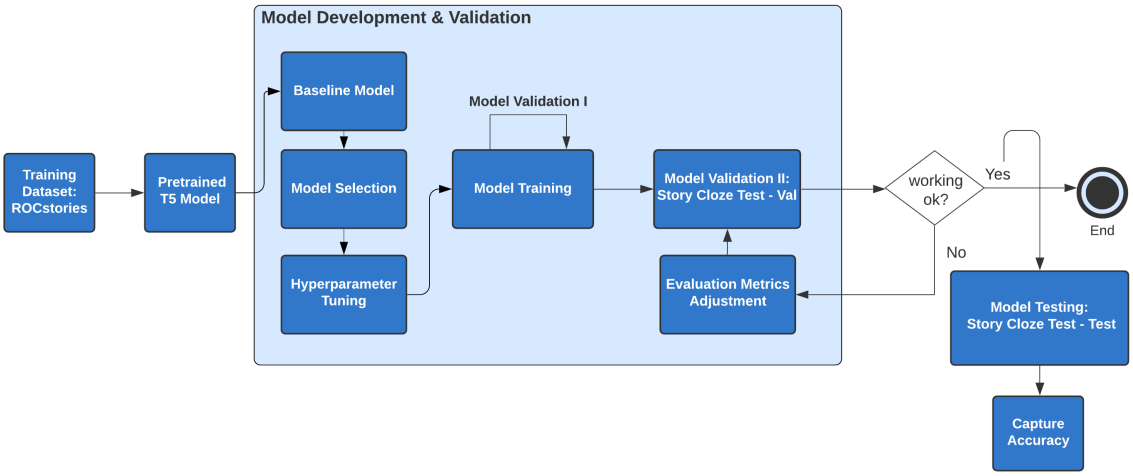


Figure 1: Pipeline

dom, backward, nearest-ending, language model and reached 67.2% accuracy.

The leaderboard of SCT v1.5 indicates that the gap between scores is quite big, due to applying different approaches. Our research is inspired by generating a generalized model to predict the next sentence under unsupervised learning (Radford, A., Narasimhan, K. , 2018)[12], where the authors demonstrated the key component of natural language understanding. Our research is innovative because none of the previous researchers has used a fine-tuned T5 model on ROC Corpus, and applied similarity score and sentiment score to test on the Story Cloze Test.

3 Methods

3.1 T5 Model

In this research, we took an unsupervised approach and used the simpleT5 (Roy, 2021)[15] model, which is pre-training on a large scale of corpus, and then fine-tuning on ROCstories training dataset. We chose T5-base model and took the following actions: set smaller batch size (batch size = 8); changed greedy search to beam search (number of beam = 10); randomly picked the next word (do_sample = True); activate Top_K sampling (top_k = 50); generated 5 outputs for each story (number of sequence = 5).

3.2 Two-Layer Model Validation

Since our training data does not have ending 1 and ending 2 options, our output is the sentence 5 prediction in free text format. Therefore, we designed two layers of validation (see Figure 1). In the first layer, the validation data was 20% randomly sampled from the training data and we used Binary Cross Entropy Loss to evaluate the performance of the text generation. In the second layer, the validation data was from the Story Cloze Test and we used the Simi-Senti score between two ending options and output to vote for the prediction and measured the accuracy by comparing the prediction with the right ending.

3.3 Evaluation Metrics Adjustment

3.3.1 Similarity Score

After obtaining the output sentences from the fine-tuned T5 model, we used two different ways to measure sentence similarity. We measured the similarity between output sentence and Ending 1 and also between output sentence and Ending 2 and then voted for the sentence that had the greater sentence similarity score.

1) **T5 Sentence Similarity.**[11] We used the pre-trained T5 model to measure sentence similarity with the prefix "stsb sentence 1: ..., sentence 2: ...'. The embedding is from the Pre-trained T5 base model.

2) **Universal Sentence Encoder.** (Cer, Yang, et al. 2018)[3]. We used the Universal Sentence Encoder large model to transfer the word text to embeddings, and then compared the similarity between output and Ending 1 and Ending 2.

3.3.2 Sentiment Score

In the Story Cloze Test, the right ending usually goes with the flow of the previous 4 sentences,

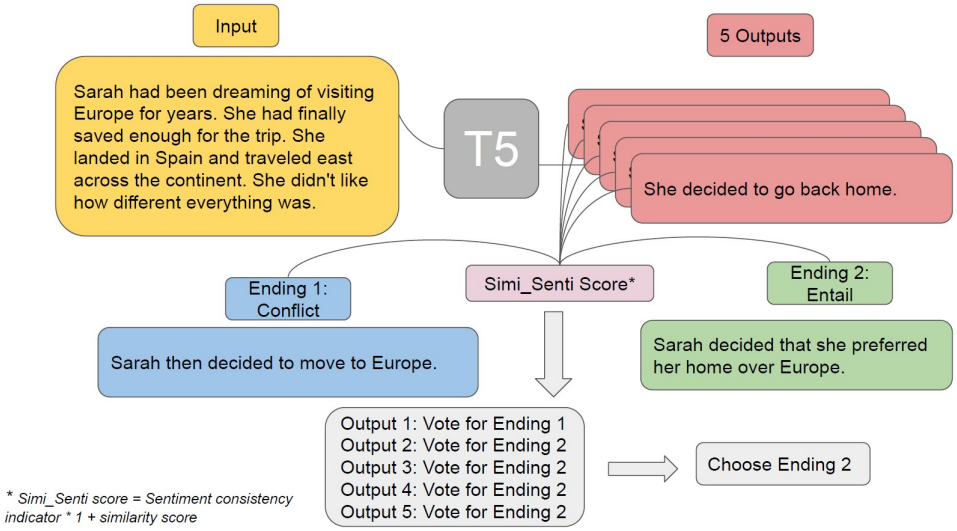


Figure 2: One Example of Improved Model Flow

which means that the right ending shares the similar sentiment with the input 4 sentences. Other than similarity scores, we used two ways to generate the sentiment score of our model output, ending 1, ending 2.

1) **VADER sentiment analysis.** VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model which is specifically attuned to the sentiments expressed in social media. It is sensitive to polarity (positive/negative) and intensity (strength) of emotion (Hutto, C.J. and Gilbert, Eric, 2015)[6].

2) **Flair - NLP sentiment analysis.**[1] Flair is a sentiment classifier model which pretrained on IMDb movies reviews and based on a character-level LSTM neural network. Flair takes the whole sentence into account and outputs "positive" or "negative" to label the sentence. This model has reached the state of arts in various datasets, which also provided us the best result for our SCT dataset.

3.3.3 Simi-Senti Score

To evaluate our model performance in the second evaluation layer, we firstly got the sentiment classes(positive, negative, neutral) for each set of output, ending 1 and ending 2. If the sentiment class matched between output and endings, we created a dummy variable to indicate that the ending sentiment is consistent with the output. Next, we calculated the similarity scores between output and two endings. We measured the similarity between output sentence and Ending 1 and also between output sentence and Ending 2. We used T5 similarity score as our main method and used Universal Sentence Encoder when the T5 similarity score is 0.0.

The distribution of the difference between two similarity scores (output and ending 1 similarity score; output and ending 2 similarity score) is likely following the normal distribution and more than half scores fall into [-1, 1]. (see Appendix C) Therefore, we decided to set weight = 1 and the Simi-Senti score calculation is shown below:

Simi-Senti Score = Sentiment Consistency Indicator * 1 + Similarity Score

Since we also experimented with generating 5 outputs for each story, we got 5 Semi-Senti scores for each ending. However, we supposed to pick one ending for each story. Therefore, we compared 5 Semi-Senti scores for ending1 and 5 Simi-Senti scores for ending 2 and each pair of the Simi-Senti scores could contribute one vote. Then we decided the final pick for the results which have the highest vote (see Figure 2).

4 Experiment

4.1 SimpleT5 baseline

For the baseline model, we used the SimpleT5 model. It takes a large-scale pretrained text-to-text T5-Small model, which has 60 million parameters[13]. Our model architecture was built on PyTorch-lightning and Transformers. For the ROC story dataset, we formatted the model input as a small paragraph, which concatenated the first 4 sentences ("sentence 1 + sentence 2 + sentence 3 + sentence 4") and the corresponding output as the

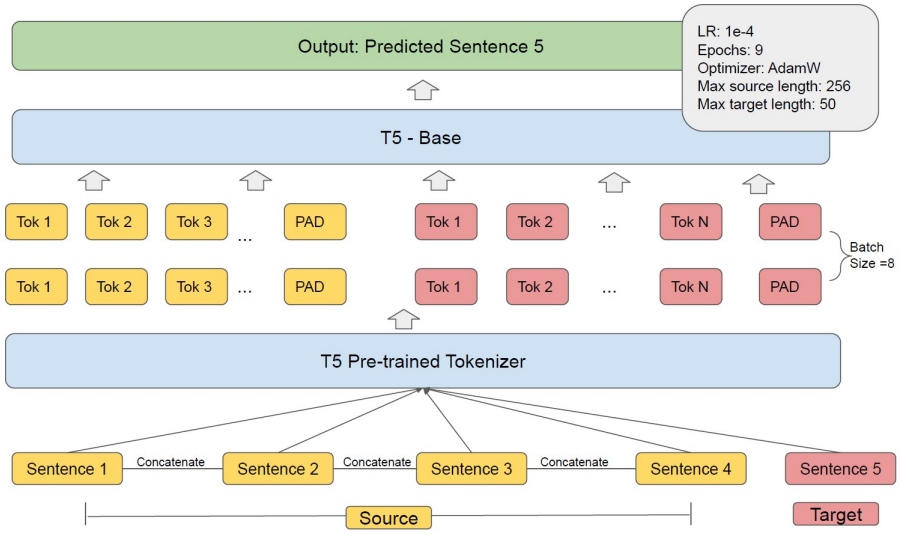


Figure 3: Model Architect

5th sentence. We then further processed the dataset as shown in Figure 3.

We fine-tuned our models in AWS by creating a G-2xlarge instance with Nvidia deep learning AMI and conducted limited parameter tuning (batch size: 8,16; learning rate: 1e-4, 1e-5, 5e-4; precision: 16, 32) to find the model that had the best performance. We used AdamW optimizer to adjust weight decay and learning rate separately. After using the best model to make the prediction on the Story Cloze dataset, we validated the model performance by using Simi-Senti score to get the accuracy score.

4.2 Story Ending Prediction Task

Firstly, we fine-tuned a pre-trained T5-small model on the ROCstories dataset. We studied some recent research papers to conduct hyperparameter tuning in order to identify the fittest model for this story ending prediction task. In addition to the T5-small model, we experimented with the T5-base model since the training dataset size is over 50,000. The T5-base model has 220 million parameters[13], which could understand the sentence complexity better. See Appendix D for the prediction result examples from T5-base and T5-small models.

In our baseline model, we only generated one ending for each story. In fact, one story could have multiple possible endings. In order to add some variance to the final output, we tried to generate 5 outputs for each story by setting Number of Return Sequence = 5 Beam Size = 10. See Appendix E for the prediction results examples for 5 outputs from the T5-base model.

5 Experiment Result

Fine-tuning T5 on the ROCstories training dataset can be challenging. It is open-ended language generation, which requires the output to be a complete sentence, have non-repeated words, and carry the logic from the input 4-sentence story[17]. In the first iteration of our T5 fine-tuning, the output sentences had no subject and included numerous repeat words. In the end, we applied simpleT5 (Roy, 2021)[15] to our training dataset instead, which generated reasonable and completed next event sentences.

Model	Epoch	Val Accuracy
T5-Small	9	64.9%
T5-Base	9	71.4%

5.1 Model Selection

After 9 epochs of fine tuning on ROCstories dataset, T5-base achieved lower train loss and higher accuracy (Table 5) on Story Cloze Test. The train loss of the T5-base model dropped more quickly than the T5-small model, reaching 1.44 loss at epoch 9 and 41.7% less than the T5-small at the same epoch. (Appendix F) The test accuracy improves 10.02%.

5.2 Parameter Tuning and Evaluation Metrics Adjustment Result

We experimented batch size 8 and batch size 16 with running 3 epochs, batch size 8 gave slightly lower val loss but took more time to run.(Appendix G) Compared to the baseline model, adding the

Model	Evaluation	Val Accuracy	% increase
T5-base 1 output	T5 Similarity Score	71.4%	
T5-base 5 outputs	T5 Similarity Score	72.3%	1.1%
T5-base 1 outputs	Simi-Senti Score	75.1%	5.2%
T5-base 5 outputs	Simi-Senti Score	76.1%	6.6%

Table 1: Parameter Tuning and Evaluation Metrics Adjustment Result

number of sequences to 5 improved the model accuracy by 1.1 %. We trained the model to output 5 endings and applied our Simi-Senti score to the model, which improved model performance by 6.6%, and reached a final accuracy of 76.1% (Table 1).

6 Error Analysis

6.1 Struggles to infer the correct ending

T5 similarity We compared the difference between similarity scores between output and two endings. We plotted the T5 similarity scores between the model output and two endings (Figure 4). It indicated that most of the time, the output was not similar to any of the ending. It was one of the biggest challenges of this task, as story ending predictions could be various. Although T5 sentence similarity outputs the float number, ultimately, the T5 model treats it as a Text-to-text problem, and transfers the float number as a string to predict the string. That is why it was not a continuous scale as displayed in Figure 5. It also meant that we needed a more dedicated method to test the sentence similarity relationship, which led us to the Universal Sentence Encoder.

Universal Sentence Encoder For the validation set, we plotted the distribution of similarity score from Universal Sentence Encoder (USE) (Figure 6). The USE scores were displayed as slightly right skewed normal distributions. Most of the similarity scores are gathered from 0.2 to 0.6, which indicates most of the cases, outputs are somewhat similar to two endings in the perspectives of the inner product of two sentence vectors. However, the same issue with T5 similarity score, we observed that the difference between two similarity scores are very close, and most of them between - 0.2 and 0.2 (Figure 7), especially in the wrong prediction set, 75% of cases that the difference between 2 similarity scores are very close.

Compared VADER and Flair two ways of doing sentiment analysis, VADER sentiment analysis is mainly based on a dictionary which maps lexical features to sentiment scores, for example, “Happy”, “Joy” are the words with positive sentiment, while “Sad” “Angry” are the words with negative sentiment. VADER calculated the sentence sentiment score based on words rather than context. In our case, 28% of sentences have zero sentiment score, which is not helpful to predict the right ending. Therefore, we introduced another pre-trained NLP model Flair.

After applying the Simi-Senti score, we significantly reduced the number of errors from 449 to 375. Among these errors, adding sentiment scores helped to capture the correct emotions that went with the flow of the stories. For the next steps, it could be helpful to use the Flair model to study the sentiment of input sentences, or the fourth input sentence, add into the weighted similarity score, which probably is helpful to find the correct ending.

6.2 Next Sentence Prediction

Overall, the fine-tuned T5 model generated reasonable endings for the input sentences. The experiment results showed that outputting 5 sentences only improved the model slightly. After diving deeper at the examples, we found that the 5 outputs are very similar as each other, with most of the time sentences only differed by the time tense or the punctuation (Appendix E). For next steps, we can also add randomness into the prediction, for example, increase the temperature, or increase the number of beams when we generate the sentence.

The output of our fine-tuned T5 model is inferred from the input sentences, while the next sentence prediction oftentimes needs to be inferred from other knowledge or common senses. One observation is that in the predicted wrong examples, the wrong ending has the key component in the input sentences (such as “airplane” or “party”), while the right ending describes the person’s feeling or describes next events which are not mentioned in the input sentences. The output shared the same words which got a much higher similarity score with the

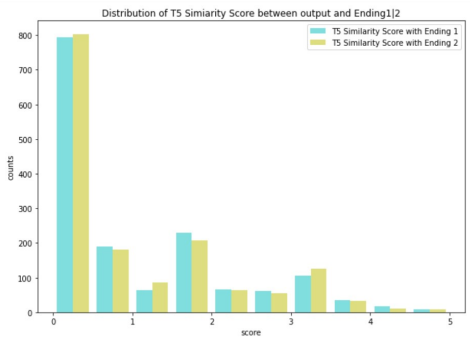


Figure 4: Distribution of T5 similarity scores

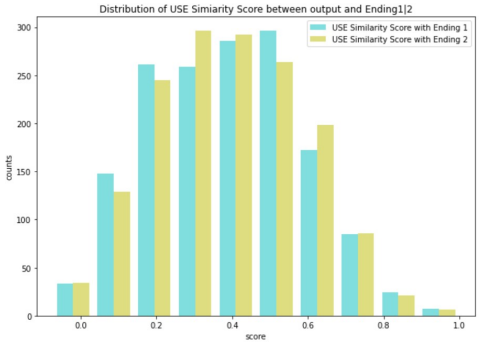


Figure 6: Distribution of USE similarity scores

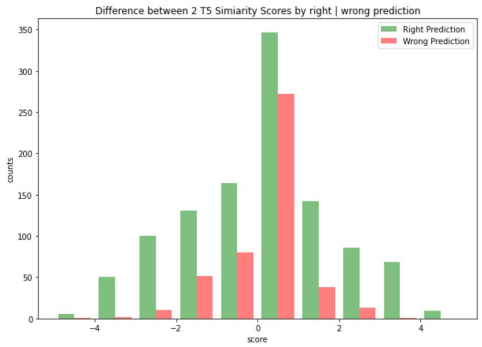


Figure 5: Distribution of the difference between T5 similarity scores

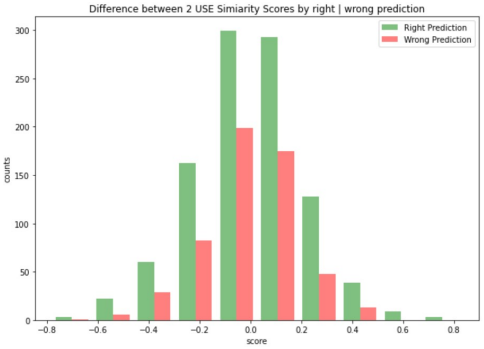


Figure 7: Distribution of the difference between USE similarity scores

wrong answer. For next steps, we can introduce common sense and extra knowledge into the model and have the model think outside of the box and generate different variations of next sentences as story endings[18].

7 Conclusion and Future Work

In this paper, we presented a training framework with the fine-tuned T5 model on ROC story corpus, along with similarity and sentiment score, which can not only generate logical and reasonable predictions for any input unlabeled text at a large scale, but also associate supervised tasks to infer the right ending for Story Cloze Test.

Our study explores two ways to improve the baseline: output more sentences from T5 model; apply sentiment score at inference. Our result indicates that applying sentiment score at inference helped the most while output more sentences only improved accuracy very limitedly. After adding the 5 outputs, we improved the accuracy by 1.1% and reached 72.2% accuracy on the validation set. After adding the Simi-Senti scores, the baseline has been improved by 5.2% and reached 75.10% accuracy. After adding the 5 outputs, and Simi-Senti scores, we reached an accuracy of 76.10%, which has increased 6.6% compared to baseline.

Future work for this study can be done through separating the input 4 sentences into "beginning", "middle", and then using both of them to predict the ending (Liu, Zhang, H., Jiang, S., Yu, D. 2019)[8]. From the linguistic perspective, the story tends to have conflict in the middle and with the contextual embeddings, the model can be trained to better understand the twists of plot, thus improving the natural language understanding[16].

Our work is publicly available for continued research and further development on this task.

Acknowledgments

We wish to express our deepest gratitude to Joachim Rahmfeld from UC Berkeley School of Information, and thank all the people who offered support along the process of completion of this research.

References

[1] Rodrigo Agerri et al. *Give your Text Representation Models some Love: the Case for Basque*. 2020. arXiv: 2004.00033 [cs.CL].

[2] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[3] Daniel Cer et al. *Universal Sentence Encoder*. 2018. arXiv: 1803.11175 [cs.CL].

[4] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[5] Yanai Elazar et al. *How Large Are Lions? Inducing Distributions over Quantitative Attributes*. 2019. arXiv: 1906.01327 [cs.CL].

[6] C.J. Hutto and Eric Gilbert. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text". In: Jan. 2015.

[7] Zhongyang Li, Xiao Ding, and Ting Liu. *Story Ending Prediction by Transferable BERT*. 2019. arXiv: 1905.07504 [cs.CL].

[8] Chunhua Liu et al. *DEMN: Distilled-Exposition Enhanced Matching Network for Story Comprehension*. 2019. arXiv: 1901.02252 [cs.CL].

[9] Lara J. Martin et al. *Event Representations for Automated Story Generation with Deep Neural Nets*. 2017. arXiv: 1706.01331 [cs.CL].

[10] Nasrin Mostafazadeh et al. *A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories*. 2016. arXiv: 1604.01696 [cs.CL].

[11] Jianmo Ni et al. *Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models*. 2021. arXiv: 2108.08877 [cs.CL].

[12] Alec Radford and Karthik Narasimhan. "Improving Language Understanding by Generative Pre-Training". In: 2018.

[13] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer". In: 1910.10683. arXiv, 2019.

[14] Melissa Roemmele et al. "An RNN-based Binary Classifier for the Story Cloze Test". In: *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 74–80. DOI: 10.18653/v1/W17-0911. URL: https://aclanthology.org/W17-0911.

[15] Shivanand Roy. "simpleT5 — Train T5 Models in Just 3 Lines of Code". In: 2021.

[16] Roy Schwartz et al. *The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task*. 2017. arXiv: 1702.01841 [cs.CL].

[17] Mingyue Shang et al. *Find a Reasonable Ending for Stories: Does Logic Relation Help the Story Cloze Test?* 2018. arXiv: 1812.05411 [cs.CL].

[18] Jiangnan Xia, Chen Wu, and Ming Yan. "Incorporating Relation Knowledge into Commonsense Reading Comprehension with Multi-task Learning". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Nov. 2019). DOI: 10.1145/3357384.3358165. URL: http://dx.doi.org/10.1145/3357384.3358165.

Appendix

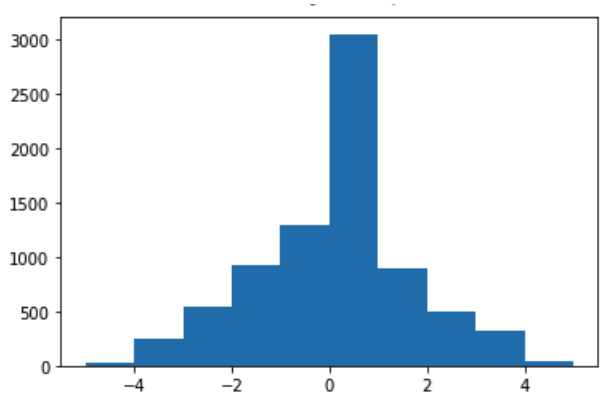
A ROCstories Sample Data

storyid	storytitle	sentence1	sentence2	sentence3	sentence4	sentence5
8bbe6d11-1e2e-413c-bf81-eaea05f4f1bd	David Drops the Weight	David noticed he had put on a lot of weight recently.	He examined his habits to try and figure out the reason.	He realized he'd been eating too much fast food lately.	He stopped going to burger places and started a vegetarian diet.	After a few weeks, he started to feel much better.

B Story Cloze Test Sample Data

InputStoryid	Input Sentence1	Input Sentence2	Input Sentence3	Input Sentence4	Ending 1	Ending 2	Right Ending
bff9f820-9605-4875-b9af-fe6f14d04256	Laverne needs to prepare something for her friend's party.	She decides to bake a batch of brownies.	She chooses a recipe and follows it closely.	Laverne tests one of the brownies to make sure it is delicious.	The brownies are so delicious Laverne eats two of them.	Laverne doesn't go to her friend's party.	1

C Histogram of Similarity Score Difference



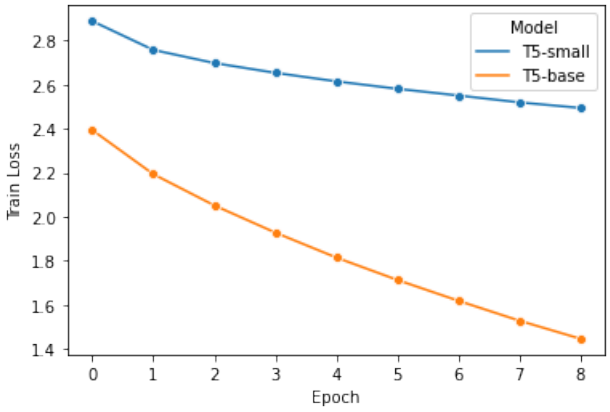
D T5-small VS. T5-base Model Output

InputStory	Ending 1	Ending 2	Right Ending	Output T5-small	Output T5-base
Laverne needs to prepare something for her friend's party. She decides to bake a batch of brownies. She chooses a recipe and follows it closely. Laverne tests one of the brownies to make sure it is delicious.	The brownies are so delicious Laverne eats two of them.	Laverne doesn't go to her friend's party.	1	Laverne is happy to have prepared a batch of brownies.	Laverne is happy to have prepared a batch of brownies.

E 5 outputs example with T5-base Model

InputStory	Output
Laverne needs to prepare something for her friend's party. She decides to bake a batch of brownies. She chooses a recipe and follows it closely. Laverne tests one of the brownies to make sure it is delicious.	1. Laverne is happy that she prepared something for her friend. 2. Laverne is happy she prepared something for her friend's party. 3. Laverne is happy that she prepared something for the party. 4. She is glad that she prepared something for her friend's party. 5. She is happy that she prepared something for her friend's party.

F Train Loss Trend Over Epochs



G Parameter Tuning: Batch Size

Model	Epoch	Batch Size	Train Loss	Val Loss
T5-base	3	8	2.3952, 2.1952, 2.0276	2.2
T5-base	3	16	2.3936, 2.1936, 2.0518	2.3

H Leaderboard Result

Results

#	User	Entries	Date of Last Entry	PercentageScore ▲
1	colormeblue1013	31	05/26/21	0.934437 (1)
2	DecstionBack	1	04/21/20	0.903246 (2)
3	malkin	2	06/29/21	0.856779 (3)
4	verbs_are_all_you_need	5	07/17/21	0.790084 (4)
5	fengyaoluo	12	12/02/21	0.745385 (5)
6	daphnei	3	04/23/20	0.721197 (6)
7	yutongl	7	02/13/20	0.709102 (7)
8	amitadk	2	11/21/20	0.515595 (8)
9	ROCNLP	1	01/16/20	0.504137 (9)
10	aamulualem	4	11/21/20	0.493316 (10)