Commonsense Knowledge Transfer from Authored to Prompt-based Generated Short Stories

Avi Bleiweiss
BShalem Research
Sunnyvale, CA, USA
avibleiweiss@bshalem.onmicrosoft.com

Abstract

In recent years, commonsense knowledge and reasoning have witnessed the emergence of a large body of research directed toward integrating commonsense into language models. Compared to encoding conventional symbolic knowledge, incorporating textual commonsense descriptions is significantly more challenging and computationally expensive. Owing to its open-ended nature, the task of training a generative story model requires both semantic and commonsense understanding. In this paper, we used both human-authored and generated short stories, and extracted from each collection commonsense data that represent sentences as subject-verb-object tuples. We finetuned our pretrained model on the commonsense authored data and followed with adaptation to the domain of commonsense story generation. To alleviate negative transfer, we applied regularized KL divergence between predicted and target tuple objects. Our quantitative analyses review the quality of commonsense knowledge transfer for out-domain genre-directed story generation compared to a baseline of cloze reading test targets.

1 Introduction

Recent advances in large pretrained language models have pushed machines closer to human-like understanding level. However, automatically assessing the quality of a generated open-ended story remains an elusive goal. The natural language generation (NLG) community largely concurs that reference based lexical overlap and relatedness of contextualized embeddings are only loosely correlated with human judgment. In OpenMEVA, Guan et al. (2021) conducted an exhaustive analysis on existing metrics and proved their lack of generalization and robustness. To improve correlation with human evaluation, UNION (Guan and Huang, 2020) proposed an unreferenced metric to learn negative sampling of generated stories, and produced a coherent score that signifies the probability to identify with a human-curated story. Complied with both machine-generated and humanly-authored stories, StoryER (Chen et al., 2022) is an evaluation metric that consists of ranking, rating, and reasoning tasks to simulate a broader extent of human preference when assessing the story quality. In this paper, we offer a more intuitive correlation with human judgment by exploring commonsense knowledge transfer from a neural model finetuned on the domain of human-authored stories and reasoned on open-ended generated stories.

Subject	Verb	Object
he	never used	claws
mom	said	never be ashamed
he	‘ll pop	ball
you	must ‘ve found	something
father	not has given	job

Table 1: A handful of svo tuples extracted from both humanly authored and generated short stories.

Automatic story evaluation presents a significant challenge for natural language models because it requires both semantic and commonsense reasoning. Commonsense knowledge graphs provide a structured way of representing a commonsense concept, which consists of a head node, a tail node, and a relation edge. The nodes in the rendition of a commonsense knowledge graph (CSKG) are normally represented by free-form short text — a word or phrase. Building knowledge graphs about each sentence in the story entails the training of a model on natural language tuples represented in an (s, v, o) base form, where s is a phrase subject, v is a verb that constitutes a relation, and o is a phrasal object. In Table 1, we show a handful examples of the generalized svo tuples.

Our contribution includes: (1) a dataset comprising high-quality human-authored children stories, and open-ended generated short stories from genre-directed prompts, (2) finetuning on svo commonsense data using KL divergence between two probability distributions to lessen negative knowledge transfer, and (3) analyses of commonsense knowledge transfer from authored short stories to both generated short story and extremely short story domains, respectively.

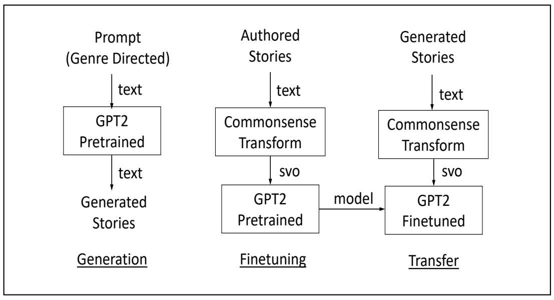

Figure 1: Breakdown of our commonsense knowledge transfer system that comprises a textual prompt-based story generation, finetuning on commonsense human-authored stories, and commonsense knowledge transfer to the domain of commonsense generated stories.

2 Background

Exposed to abundant amount of data, large pretrained language models are considered knowledge bases on their own, however, to efficiently embody the knowledge into downstream NLG tasks is often non-intuitive.

Zheng et al. (2023) proposed a finetuning technique that mines pretrained knowledge in large language models. They frame plain finetuning into a causal graph and relate catastrophic forgetting to the vanishing of causal effects in pretrained data. Their method differentiates the strength of knowledge preservation by selecting a subset of most nearest neighbors for each sample to alleviate negative transfer. Running almost exclusively on commonsense QA datasets, empirical evidence suggest their method outperforms a handful of baselines.

Zhou et al. (2023) proposed a framework to transfer the commonsense knowledge stored in a neural commonsense knowledge model to a general-purpose pretrained language model on a large-scale and diverse text corpus. Notably their source (ATOMIC; Hwang et al., 2021) and target (T5; Raffel et al., 2020) models are heterogeneous, and their analyses prove to outperform models that exploit either symbolic knowledge graphs or texts alone. However, consuming over 1B network parameters combined poses a system resource constraint and less than optimal computational complexity.

To improve computational efficiency for integrating textual commonsense description in language models, Cui and Chen (2023) proposed the split of training examples with similar descriptions. Their model achieved comparable efficacy to a baseline OK-Transformer (Cui and Chen, 2022) that randomly partitions examples into batches. On average, it reduces the time cost for knowledge encoding by 40% for commonsense reasoning tasks. Nonetheless, this optimization might not generalize to a generative pretrained transformer (GPT) that implies serial execution and a batch size of one.

3 Data

Our textual source data is split between human-authored and generated short stories, from which we extract commonsense tuples in the subject-verb-object (svo) form.[1]

Human-Authored Stories We obtained unicode encoding of the human-authored literature text from Project Gutenberg, and carried out our work on thirty seven children books from the FreeChildrenStories.com collection by Daniel Enrico (2009). [2] Table 2 provides statistical distribution of typical text properties. These narratives total about 60K tokens and average over 1.6K words per story. We also post for the literary set the Flesch Reading Ease (FRE) score that identifies a difficulty level range from fairly difficult to very easy.

Item	Min	Max	Mean	STD
Tokens	518	3,524	1,638.8	753.9
Sentences	10	78	35.7	14.1
Characters	547	3,765	1,730.9	794.9
FRE	48.8	104.5	74.9	12.5

Table 2: Statistical distributions of textual surface properties for our authored stories.

Generated Stories We scraped ten short story prompts for each of the following genres: fiction, comedy, fantasy, mystery, and romance. [3] From which we generated a total of fifty short stories using GPT2. We thus had ten stories for each category, with an average of close to 4,000 words per story and a total of about 200K tokens (Table 3). In contrast to authored stories, the FRE scores for generated content point to a fairly difficult reading level. A sample of a textual fiction-based prompt

[1] The generated story dataset is available at https://github.com/bshalem/sge

[2] https://www.gutenberg.org/cache/epub/29762/pg29762.txt

[3] https://www.tckpublishing.com/short-story-prompts/

followed by three generated continuation sentences are shown in Table 4.

Item	Min	Max	Mean	STD
Tokens	3,507	4,192	3,913.1	150.6
Sentences	30	69	41.5	7.3
Characters	3,710	4,322	4,055.3	137.8
FRE	1.55	65.3	28.1	14.4

Table 3: Statistical distributions of textual surface properties for our generated stories.

prompt: A college-age son struggles to win his father's approval, especially when he quits college to pursue his dreams of becoming an artist.

continuation: It would sound good on her head if some day they didn't meet him at The gallery!
continuation: His love interest also takes over and runs a art store for his mother while we talk about finding money here;
continuation: however you can see how much I hate seeing what happens outside this window just walking around my door after she left?

Table 4: A textual surface sequence of an initial prompt following by generated continuations in a fiction story.

Commonsense Base We used the spaCy library to identify a sentence structure from a parsed dependency tree and extracted an ordered sequence of subject-verb-object triples from a document or sentence. In Table 5, we provide statistical distribution of commonsense tuples for svo transformed humanly-authored and generated short stories; tuples total 648 and 2,197, respectively, representing both main and subordinate clauses.

Stories	Min	Max	Mean	STD	Total
Human	3	39	17.5	8.0	648
Generated	30	72	43.9	7.8	2,197

Table 5: Statistical distributions of commonsense tuples across our human authored and generated stories.

4 Method

In Figure 1, we review our three-step evaluation process: (i) generation, (ii) finetuning, and (iii) transfer. Computationally, at each stage we maintain both a plain textual surface and a commonsense svo representations of the data. Our goal is to automatically assess the closeness of open-ended story generation to human-authored stories. We conduct this evaluation entirely in the three-dimensional commonsense space as outlined by the svo format.

Generation Using a pretrained GPT2 model, we generate short stories from a genre-specific prompt. In our proposed framework for commonsense knowledge transfer we used the transformer-based (Vaswani et al., 2017) GPT2 language model (Radford et al., 2019). The model was finetuned on a random sample of sentences from the BookCorpus dataset (Zhu et al., 2015) that was used to pretrain BERT (Devlin et al., 2019). [4] In our study, this model is intended for generative genre-oriented short stories. [5] The model is fed by a genre directive that follows the beginning of sequence token (BOS) and precedes the input prompt in the form

$$< \text{BOS} >< \text{genre} > \text{a short input prompt},$$

and supports the following five narrative classes: fiction, comedy, fantasy, mystery, and romance.

Finetuning Textual human-authored stories are transformed to the commonsense svo format and feed our GPT2 model to undergo finetuning. We drew upon the finetuning approach proposed by Aghajanyan et al. (2021) that regularizes KL divergence between two probability distributions. Their method has shown to outperform standard finetuning on three downstream summarization tasks, while being computationally affordable. We used the Kullback-Leibler Divergence (KL; Kullback and Leibler, 1951) loss that computes the logarithmic difference between two probability distributions. KL divergence quantifies the dissimilarity of the prediction probability distribution from the grounded probability distribution. In our implementation, both the prediction and target probabilities of the tuple object are provided in logarithmic space for efficient computation. More formally the KL distance is defined as:

$$D_{KL}(Y||X) = \sum_{i=1}^{N} \log\left(\frac{Y_i}{X_i}\right) Y_i,$$

where N is the number of tuple samples, and X and Y are the probability distributions of the observation and reference, respectively.

[4] https://huggingface.co/datasets/bookcorpus

[5] https://huggingface.co/aspis/gpt2-genre-story-generation

Transfer In this stage, we transfer commonsense knowledge from authored to generated story domains and validate the quality of transfer learning using a self-sufficient metric. In commonsense mode we feed the GPT2 language model with the concatenation of a subject-verb pair and use the object of the knowledge tuple as the reference target. Formally, we used the colon notation $s_{1:n} = (s_1, \ldots, s_n)$, $v_{1:m} = (v_1, \ldots, v_m)$, and $o_{1:l} = (o_1, \ldots, o_l)$ to denote an n-token subject, m-token verb, and l-token object, respectively. The input concatenation $[s_{1:n}; v_{1:m}]$ has thus the dimensionality of $(n+m)$. Our GPT2 model is finetuned to predict object tokens from this concatenation. To evaluate the quality of commonsense transfer from in-domain human-authored to out-domain generated short stories we used Sentence Transformers (SBERT; Reimers and Gurevych, 2019). [6] SBERT computes contextualized embedding similarity between the predicted object and reference target.

5 Evaluation

In our experiments we analyzed the performance of commonsense knowledge transfer from human curated to generated short stories. To extract subject-verb-object tuples from a textual surface we used spaCy. [7]

Genre	Perplexity	Burstiness	FRE
Fiction	90.5	14.1	**30.8**
Comedy	87.7	16.9	22.4
Fantasy	91.5	16.7	25.7
Mystery	**81.7**	17.1	23.6
Romance	87.3	**19.4**	15.7

Table 6: Genre-specific average perplexity, burstiness, and FRE across our generated stories.

Perplexity and Burstiness Generative language models are measured by the ability to generate coherent and contextually relevant text. To this end, perplexity and burstiness are closely related concepts in text generation. Perplexity measures the overall uncertainty of a language model, while burstiness examines the local distribution of words or phrases. A lower perplexity indicates better prediction accuracy, and a lower burstiness suggests a more monotonous plot style. In Table 6, we show genre-specific average perplexity and burstiness across our machine-made stories. Mystery content leads favorably on perplexity and romance presents the more vibrant matter. Drawing on FRE scores validates generative fiction to be the easiest to read.

Continuation Relevance We assessed quantitatively how relevant are each of the generated continuations to the story context, the prompt. The metric we chose was cosine-similarity computed between continuation and prompt embeddings that were produced by SBERT, and the result scaled to $[0.0, 1.0]$. Regardless of the story genre, we anticipated a relevance decline the farther apart the continuation from the prompt in the plot timeline.

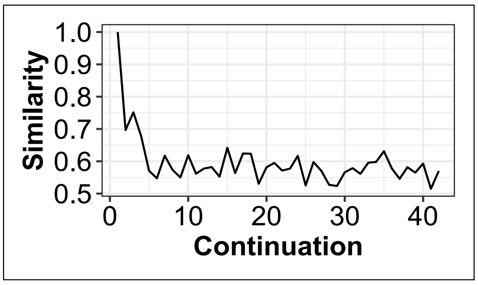

Figure 2: Continuation-context similarity as we advance in the story-plot timeline.

In Figure 2, we show the rendered continuation-context relevance for a fictional short story that comprises a total of 42 continuations, including the leading prompt. At first, similarity drops precipitously for the first handful of continuations and then it levels off and declines more moderately. This behavior appears rather consistent in all genres and suggests a model generation constraint. In Table 7, we show complementary statistical distribution of continuation-context relevance across our generated story collection. We post the second largest similarity element, 0.87, as the largest is 1.0 for matching the story guiding prompt.

Relevance	Min	Max	Mean	STD
Similarity	0.52	0.87	0.57	0.08

Table 7: Statistical distribution of continuation-context relevance across all our generated stories.

Training Setup The GPT2 model was finetuned on our commonsense data we derived from the human-authored children narratives. Our model was trained for ten epochs using the KL divergence loss, the AdamW optimizer, an initial learning rate of 1e−5, and a batch size of one due to the GPT2 generation nature of one token per iteration. Fine-

[6] https://huggingface.co/sentence-transformers
[7] https://spacy.io/

tuned on a sample of the BookCorpus for genre-based short story generation, our GPT2 checkpoint has an underlying neural network of over 124 million parameters.

In figure 3, we show comparative loss behavior for ten epochs between KL divergence and Wasserstein distance metrics. Both are reasonably spiky, however, KL appears to start its descent toward converging earlier compared to Wasserstein.

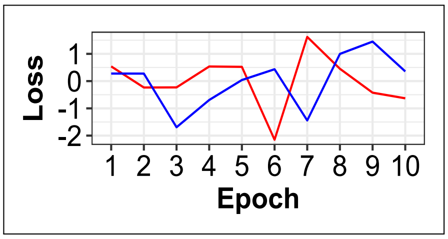

Figure 3: Correlation between KL divergence (red) and Wasserstein distance (blue) loss functions.

Story Domain Distilling Our finetuning process effectively incorporates commonsense of the authored stories in our GPT2 model. Using the svo version of our open-ended generated short-stories introduces an out-domain commonsense with respect to the in-domain human-authored stories to enable analyze the quality of knowledge transfer. We report average embedding similarity across narrative genres or the entire corpus. Known to be better suited for transfer learning than BERT, our SBERT model encodes for each svo tuple the sentence embeddings of a predicted and a target object pair of which we compute cosine similarity scores. The SBERT model renders about 82 million trained parameters.

Genre	Min	Max	Mean	STD
Fiction	0.155	**0.231**	0.184	0.020
Comedy	0.160	0.213	0.184	0.019
Fantasy	0.159	0.204	0.181	0.014
Mystery	0.154	0.214	0.188	0.019
Romance	0.174	0.206	**0.190**	0.011

Table 8: Genre-specific statistical distributions of commonsense knowledge transfer rates.

In Table 8, we review the performance of our commonsense knowledge transfer. Rates suggest the correlation between open-ended generated stories and hand-written stories. On average, the romance genre is leading with a 0.19 similarity score, while the fiction genre scored the highest rate of 0.23. Overall, results appear fairly commensurate given the size of our children short-story corpus. We expected children stories destined to young adults to mainly fall in one of fiction, mystery, and fantasy genres and affect our leading distilled scores proportionally. Evidently the highest genre transfer rates shown in Table 8 are ranked $(0.23, 0.21, 0.20)$ for fiction, mystery, and fantasy, respectively. In Figure 4, we provide complementary visualization of density estimates for transfer rate distribution.

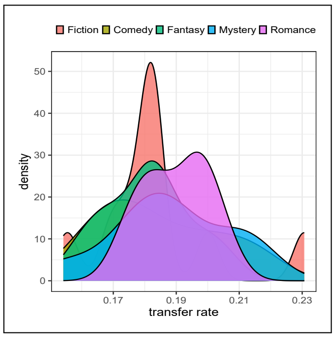

Figure 4: Density estimates across narrative genres for commonsense distribution of knowledge transfer rates.

Extremely Short Stories As a baseline for our experiments, we sought after a dataset that consists of extremely short stories, and found the Story Cloze Test (SCT-v1.5; Sharma et al., 2018) corpus a useful match. SCT-v1.5 has 1,571 examples, each comprises a four sentence story context along with a single-phrase ending pair. The task that SCT prescribes for the underlying model is to correctly choose the ending given the context. However, in our evaluation, we were interested in the transfer rate quality of commonsense knowledge from human authored stories to the SCT domain. In the first processing step we transformed the SCT story context from a textual surface to the commonsense svo format and followed with an inference task. Table 9 shows statistical performance distribution of commonsense transfer to the SCT domain with a mean and maximal scores of 0.16 and 0.45 compared to 0.19 and 0.23 for the generated data, respectively.

SCT-v1.5	Min	Max	Mean	STD
Context	0.003	0.449	0.164	0.052

Table 9: Statistical performance distribution of commonsense knowledge transfer for SCT context.

External Baseline Comparison To the extent of our knowledge based on public research, we are the first to offer separable genre evaluation of generated short stories. This makes it challenging to perform even-handed quality comparison with external baselines, each using different corpora, evaluation model, and metrics.

System	Data	Model	Rate
StoryER	WP	Longformer	0.09
OpenMEVA	WP	GPT2	0.16
COMET	ATOMIC	GPT	**0.27**
Ours	MultiGenre	GPT2	0.18

Table 10: Comparing quality of automatic story evaluation to external baselines.

In table 10, we list data, language models, and performance for three external baselines. StoryER (Chen et al., 2022) and OpenMEVA (Guan et al., 2021) use the WritingPrompts (WP) dataset,[8] as COMET (Hwang et al., 2021) explores ATOMIC, a general purpose commonsense knowledge resource. Checkpoints of GPT models appear to dominate the baseline selection of the Transformer model. All models are at least pretrained on a large text collection and most are finetuned on an in-domain data split. We chose a referenced metric common to all baselines and used the reported top scoring BLEU (Papineni et al., 2002)— more precisely the geometric mean of n-gram, with $n = \in (1,2,3,4)$. Our method uses multi-genre generated stories and provides quality of knowledge transfer from one story domain to another. Our transfer rate is measured as embedding similarity and in Table 10 we post the average score of all genres. After adjusting our methodology to better fit the baselines, we show COMET to lead at 0.27 BLEU as our method is second with 0.18 BLEU.

6 Discussion

Conceptually relevant to our study, Chambers and Jurafsky (2009) investigated learning unsupervised narrative schemas. Their algorithm uses coreferring arguments in chains of verbs to learn both rich narrative event structure and argument roles. A narrative schema is defined as a 2-tuple set of events and typed chains over the event slots. Understanding causal relationship between events either adds a slot to an existing chain or adds a new chain.

Our work uses a 3-tuple subject-verb-object representation of an event and computes for each svo tuple the similarity between predicted and reference objects. We further accomplish our main goal of scoring a story by averaging similarities of all the tuples that constitute a story. However, to interpret a story as a continuation set of narrative events, requires our method to evolve toward computing the relatedness between svo tuples and forming a narrative graph that links tuples based on strong similarities.

7 Conclusion

We proposed a novel and intuitive approach for automatic evaluation of machine generated short stories. In this paper, we recast the problem of correlation between generated and human curated narratives to transferring knowledge from one story domain to another. We conducted our experiments entirely in the commonsense multidimensional space, and used KL divergence for finetuning our language model to present effective integration of encoding structured commonsense. Applying commonsense reasoning to our task showed distinct genre performance behavior and scalable rates for extremely short stories.

8 Limitations

At first observation, the number of human-authored short stories appears rather small. While it is plausible to gain knowledge transfer quality by augmenting the collection, our study goals were met by showing discernible performance across narrative genres. Although KL divergence has been widely adopted for finetuning language models on summarization tasks, it is our conjecture that for more generalized generation objectives, we would like to explore other approaches, such as the Jensen-Shannon divergence, to improve upon the performance of commonsense knowledge transfer.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful suggestions and feedback.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2021. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*.

[8] https://huggingface.co/datasets/euclaise/writingprompts

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the Annual Meeting of the ACL and the International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Hong Chen, Duc Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2022. StoryER: Automatic story evaluation via ranking, rating and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1739–1753, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wanyun Cui and Xingran Chen. 2022. Enhancing natural language representation with large-scale out-of-domain commonsense. In *Findings of the Association for Computational Linguistics (ACL)*, pages 1746–1756, Dublin, Ireland. Association for Computational Linguistics.

Wanyun Cui and Xingran Chen. 2023. Free lunch for efficient textual commonsense integration in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3759–3770, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jian Guan and Minlie Huang. 2020. UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9157–9166, Online. Association for Computational Linguistics.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 6394–6407, Online. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6384–6392.

S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing and the Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the story ending biases in the story cloze test. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 752–757, Melbourne, Australia. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (neurIPS)*, volume 30. Curran Associates, Inc.

Junhao Zheng, Qianli Ma, Shengjie Qiu, Yue Wu, Peitian Ma, Junlong Liu, Huawen Feng, Xichen Shang, and Haibin Chen. 2023. Preserving commonsense knowledge from pre-trained language models via causal inference. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9155–9173, Toronto, Canada. Association for Computational Linguistics.

Wangchunshu Zhou, Ronan Le Bras, and Yejin Choi. 2023. Commonsense knowledge transfer for pre-trained language models. In *Findings of the Association for Computational Linguistics(ACL)*, pages 5946–5960, Toronto, Canada. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *International Conference on Computer Vision (ICCV)*, pages 19–27.