Tackling BabyLM challenge using transformer based approaches

Can Gözpınar
Koc University
cgozpinar18@ku.edu.tr

Aydin Ahmadi
Koc University
aahmadi22@ku.edu.tr

Abstract

Recently, Language models which are usually trained in a self-supervised manner have gained attention of the researchers. The focus of these models is to learn a better representation for words, sentences to utilize them to generate a high-quality text. Despite the advances in language modeling, the pretrained language models do not generate coherent texts when applied to downstream tasks. BabyLM challenge[1] has been presented recently which provides a limited-sized corpus inspied by the spoken and written language which is presented to children. We propose to devise a sample-efficent language model to tackle the strict-small dataset challenge.

1 Introduction

The BabyLM challenge involves a dataset of spoken and written language acquired by children during their early years. Developing an effective model to tackle this challenge presents significant hurdles. Firstly, the dataset is relatively small, making it difficult to train a language model. Secondly, most pre-trained tokenizers contain a vast number of tokens that are not useful for this dataset and will not aid in model performance. Thirdly, the language in the dataset is ordered according to a child's language acquisition, with words in the Aochildes file representing those heard by infants under one year of age, and those in the Wikipedia file being more complex, representing the language of teenagers. Additionally, sentence and passage lengths in the dataset vary significantly. These factors require careful consideration in designing a tokenizer and model architecture that can effectively handle this challenge. Overcoming these challenges in developing a language model for a limited corpus can significantly benefit the natural language processing and cognitive science communities. Our project aims to explore various methodologies for dataset preprocessing and training, as well as implementing novel techniques in our model architecture pipeline.

2 Related Work

The core of the text generation task is to model a mapping from input to output. The inputs can have various forms like graphs, tables, or multimedia. The outputs are the varied-size of the sequences of a text. The BabyLM challenge considers the inputs as sequences of words and outputs as a sequence which is generated by a language model. Various methods have been proposed to solve the text generation tasks based on RNN (Chen et al., 2020), CNN (Gehring et al., 2017), GNN (Li et al., 2020) and attention mechanism (Bahdanau et al., 2015) . However, with the advent of transformer architecture (Vaswani et al., 2017), using pretraining language models have gained the attention of researchers. The idea behind the PLM's is to train the models in large-scale corpus and then fine-tune them in a downstream task. These models can achieve a universal representation of the language which can be beneficial to be utilized in a downstream task instead of training the model from scratch. Brown et al. (Brown et al., 2020) Have particularly shown this. The trained transformer-based models vary in their usage of the transformer architecture. While architectures like T5 (Raffel et al., 2020) and Bart (Lewis et al., 2020) use a standard encoder-decoder architecture of transformers, models like GPT (Brown et al., 2020) utilize their decoder-only architecture and Bert (Devlin et al., 2019) encoder-only. Raffel et al. (Raffel et al., 2020) have had a conclusion in their paper that using the encoder-decoder architecture is more beneficial. The following part will present some recent research on using different methodologies that aims to have meaningful generated text by leveraging the pretrained language models.

Lee et al. (Lee et al., 2020) have presented a work utilizing a transformer-based sentence-level

Dataset	Domain	# Words		Proportion
		STRICT-SMALL	STRICT	
CHILDES (MacWhinney, 2000)	Child-directed speech	0.44M	4.21M	5%
British National Corpus (BNC),[1] dialogue portion	Dialogue	0.86M	8.16M	8%
Children's Book Test (Hill et al., 2016)	Children's books	0.57M	5.55M	6%
Children's Stories Text Corpus[2]	Children's books	0.34M	3.22M	3%
Standardized Project Gutenberg Corpus (Gerlach and Font-Clos, 2018)	Written English	0.99M	9.46M	10%
OpenSubtitles (Lison and Tiedemann, 2016)	Movie subtitles	3.09M	31.28M	31%
QCRI Educational Domain Corpus (QED; Abdelali et al., 2014)	Educational video subtitles	1.04M	10.24M	11%
Wikipedia[3]	Wikipedia (English)	0.99M	10.08M	10%
Simple Wikipedia[4]	Wikipedia (Simple English)	1.52M	14.66M	15%
Switchboard Dialog Act Corpus (Stolcke et al., 2000)	Dialogue	0.12M	1.18M	1%
Total	–	9.96M	98.04M	100%

Figure 1: The datasets for the STRICT and STRICT-SMALL tracks of the BabyLM Challenge. CHILDES (child-directed speech), OpenSubtitles (speech), BNC (speech), TED talks (speech), children's books (simple written language are subsets of these datasets.

tokenization model. They added trainable sentence embeddings indicating the index of each sentence to each of the word representations. They trained the sentence embedding on a BERT to have the encoded representations. Then, they used a sequence reconstructor which consists of a decoder and a pointer network to predict the sequence order after shuffling input sentences. Their results show that their model is comparable to T5. Araujo et al. (Araujo et al., 2021) presented a work that similarly uses sentence-level representations. They got inspiration from predictive coding (van Berkum et al., 2005) which is a neuroscience theory on language development. However, they predict the next future sentence on a top-down pathway with a next-sentence loss + BERT loss. They discuss that their model learns discourse-level representations and sentence relationships. Lee et al. (Lee et al., 2022) presents a concept-based curriculum learning approach to language modeling. They work a masking mechanism in a curriculum that begins with the core concepts in human language acquisition and ends with complex ones. In their approach, They first identify some of the core concepts in words and phrases to be masked by scoring concepts in a ConceptNet in a top-down manner. They gradually mask concepts related to the previous masked concepts in the consecutive stages. They show the effectiveness of their algorithms for masking in masked language models in their paper.

Ji et al. (Ji and Huang, 2021) introduces a novel latent variable model that learns a sequence of discrete latent variables from long text and utilizes them to guide the text generation process, ensuring coherence in the output. This allows the model to abstract the discourse structure of the text and generate coherent long texts with interpretable latent codes.

3 Model

We will train Transformer models on the provided BabyLM dataset from scratch. We will not be using pre-trained models to fine-tune on our task since BabyLM challange does not permit any models that is trained on different datasets to be used. We plan to experiment with different Transformer architectures with different pre-training approaches. We will experiment with different Transformer architectures (initially we plan to start with T5), initialization schemes and hyperparameters. We will be using datasets module from huggingface to load and process our dataset. Transformers module from huggingface for pre defined transformer models implementations and tokenizers. We will rely on PyTorch for its deep learning utilities and Tensor manipulations. We will clean our dataset using the BabyLM Challange's data preprocessing pipeline `https://github.com/babylm/babylm_data_preprocessing`. Since we will be training our model from scratch we expect our task to be computationally more expensive than other tasks that can leverage the benefit of fine-tuning a pre-trained language model. For the GPU resources we plan to start by using Google colab's free GPUs, and if it does not suffice we apply for student credits from AWS, and Google Cloud.

4 Experimental Setup for Your Approach

This year BabyLM challange will be accepting its first submissions as a result there are no directly applicable previous approaches to our task. Our task is interested in small scale language modeling with small dataset. Even though their tasks may differ, we are inspired by the previously published approaches which tried to train transformer models from scratch on relatively small scale datasets. Inspired by (Xu et al., 2021) we investigate the effects of *T-Fixup* (Huang et al., 2020) with the hopes of having a more stable training process on our small scale dataset. While modifying the model layers and its hyperparameters, we will refer to the findings of (Tay et al., 2022). Since, the BabyLM challange offers a baselines for some of the available models in the huggingface's Transformers module, we will start with those available models such as *T5*. Furthermore, we will try a curriculum learning like approach in which we will try to train model from increasingly harder data as it gradually learns to handle easier data first . We plan to achieve this by training our model on data that we deem more basic first such as data from childrens books (Children's stories text corpus) and children-directed speech (CHILDES), and then moving to harder to grasp dataset such as Wikipedia. We got inspired by (Lee et al., 2022) but reverted to our framework of curriculum learning due to the strict limitations imposed onto us by BabyLM Challange such as forbiding any other dataset, or integration of any model which is trained on any other dataset. For evaluating our model we will use the evaluation evaluation pipeline of the BabyLM Challange (`https://github.com/babylm/evaluation-pipeline`) along with the trivial metrics such as model losses.

4.1 Dataset

The dataset which we will use is the STRICT SMALL track of the BabyLM challenge which is shown in 1. The dataset is avalable in the Github page of the challenge.

4.2 Baseline Models

The baseline methods are presented in the evaluation pipeline Github page of the challenge. They have considered OPT-125m, RoBERTa-base, T5-base models as baslines for the challenge. They are naive baselines that are meant to provide a starting point for investigation as being said in the challenge evaluation-pipeline page. They are obtained by staying as close to the reported values in their corresponding original implemenatations as possible and by no means they are delicatel fine-tuned models that reflect the true potential they can achieve.

4.3 Evaluation Metrics

The evaluation metrics for the task is published in the evaluation pipeline page of the task. But, for the clarification they are as explained briefly below.

Anaphor agreement is a linguistic concept that refers to the grammatical agreement between a pronoun and its antecedent in terms of gender, number, and person.

Argument structure, on the other hand, is the way that a verb selects the number, type, and order of its arguments.

Binding refers to the relationship between a pronoun and its antecedent, while control raising occurs when a non-finite verb takes on the subject of its matrix clause.

Ellipsis is a linguistic concept that involves the omission of recoverable words, while filler-gap describes the relationship between a gap and a filler.

NPI licensing is the mechanism by which negative polarity items are licensed, and NPI licensing quantifiers are words that license the use of NPIs in negative contexts.

subject-verb agreement is a rule in grammar that requires the verb in a sentence to agree with its subject in terms of number and person.

5 Schedule

We have devided the subtask of our project as below. We have scheduled considering the timelines in project info slide.

1. pre-processing the data - By May 5th
2. Build pretrained models for task - By May 12th
3. Build torch based models for the task to be trained - By May 20 th
4. Build and train curriculum based approaches - By May 25 th
5. Analyze the output of the model, and do an error analysis - By May 30th
6. Loop thought the third item in the list if needed- By June 7th
7. Work on final presentation - By June 13th
8. Work on final report - By July 1st

References

Vladimir Araujo, Andrés Villa, Marcelo Mendoza, Marie-Francine Moens, and Alvaro Soto. 2021. Augmenting BERT-style models with predictive coding to improve discourse-level representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3016–3022, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2020. Few-shot NLG with pre-trained language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Online. 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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*.

Xiao Shi Huang, Felipe Perez, Jimmy Ba, and Maksims Volkovs. 2020. Improving transformer optimization through better initialization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4475–4483. PMLR.

Haozhe Ji and Minlie Huang. 2021. DiscoDVT: Generating long text with discourse-aware discrete variational transformer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4224, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haejun Lee, Drew A. Hudson, Kangwook Lee, and Christopher D. Manning. 2020. Slm: Learning a discourse language representation with sentence unshuffling. In *Conference on Empirical Methods in Natural Language Processing*.

Mingyu Lee, Jun-Hyung Park, Junho Kim, Kang-Min Kim, and SangKeun Lee. 2022. Efficient pre-training of masked language model via concept-based curriculum masking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7417–7427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Junyi Li, Siqing Li, Wayne Xin Zhao, Gaole He, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. 2020. Knowledge-enhanced personalized review generation with capsule graph neural network. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 735–744, New York, NY, USA. Association for Computing Machinery.

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(140):1–67.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2022. Scale efficiently: Insights from pre-training and fine-tuning transformers.

Jos J. A. van Berkum, Colin M. Brown, Pienie Zwitserlood, Valesca Kooijman, and Peter Hagoort. 2005. Anticipating upcoming words in discourse: evidence from erps and reading times. *Journal of experimental psychology. Learning, memory, and cognition*, 31 3:443–67.

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*, volume 30. Curran Associates, Inc.

Peng Xu, Dhruv Kumar, Wei Yang, Wenjie Zi, Keyi Tang, Chenyang Huang, Jackie Chi Kit Cheung, Simon J. D. Prince, and Yanshuai Cao. 2021. Optimizing deeper transformers on small datasets.