BoyBERTa - Experimenting on BabyBERTa for Grammar Learning and Language Understanding

Phakphum Artkaew*and Vorrapard Kumthongdee*

Department of Electrical and Computer Engineering, New York University
pa2497@nyu.edu, vk2584@nyu.edu

Abstract

Advances in Natural Language Processing (NLP) have been significantly driven by models like BERT and RoBERTa, which excel in various linguistic tasks due to their deep learning capabilities. However, these models often require extensive data and computational resources, which can limit their adaptability for specific tasks like grammar learning and nuanced language understanding. This study introduces BoyBERTa, an optimized version of the BabyBERTa model, developed to enhance grammar learning from child-directed speech. Utilizing a novel curriculum learning approach, BoyBERTa is trained on a 10-million-word corpus, progressively moving from simpler to more complex linguistic tasks. The training effectiveness is evaluated against a random order training approach using the BLiMP benchmark. Our results demonstrate that while BoyBERTa shows improved performance in certain linguistic tasks, the benefits of curriculum learning in this context are subtle, suggesting that the model's configuration and training methodology play critical roles in its overall effectiveness. BoyBERTa's performance is also compared to the baseline BabyBERTa and RoBERTa models, highlighting its efficiency and potential utility in educational and developmental language applications. This paper further discusses the implications of these findings for future NLP model development and the potential of tailored training approaches in enhancing model understanding of complex language structures. The repository can be accessed at https://github.com/PhakphumAdev/BabyBERTaMax

Introduction

In recent years, the field of Natural Language Processing (NLP) has witnessed remarkable advancements, primarily due to the development of large-scale transformer-based models [11] such as BERT (Bidirectional Encoder Representations from Transformers) [3] and RoBERTa (Robustly Optimized BERT Approach) [7]. These models have revolutionized tasks such as sentiment analysis, machine translation, and question-answering, offering unparalleled performance due to their ability to capture deep contextual information [6]. However, despite these significant strides, challenges remain in optimizing these architectures for specific tasks like grammar learning and language understanding [9].

While RoBERTa stands out for its enhanced robustness and fine-tuning capabilities over its predecessors, its design makes the model difficult to handle multiple tasks efficiently. The model requires fine-tuning for each task and lacks flexibility in handling multiple types of text generation tasks [8]. In addition, due to the enormous size of training data and the complexity of the model, RoBERTa is vulnerable to certain types of adversarial examples and overfitting to spurious correlations, which means that the model may struggle to generalize in specific contexts with data distribution shifts [10]. This leads to another approach of the pre-training Natural Language Processing model.

Instead of a complex model, BabyBERTa [5] was trained exclusively on a small dataset of child-directed speech to investigate its ability to learn grammar, aiming to understand how such models process the unique grammatical structures present in child-directed language and compares their learning against other models. By training a specialized language model like BabyBERTa on child-directed speech can yield significant gains in grammar learning, providing insights into how exposure to different types of language data affects model comprehension. [5].

This paper introduces BoyBERTa, an advanced and optimized pre-trained version of BabyBERTa, tailored to enhance grammar learning and deepen language comprehension. BoyBERTa is designed for efficiency, clarity, and flexibility in addressing diverse linguistic patterns, featuring targeted enhancements that significantly improve its capability to recognize and interpret grammatical structures. By utilizing a larger dataset and a varied training approach, BoyBERTa's effectiveness is assessed by comparing its performance with both the standard BabyBERTa models and the RoBERTa-base model across multiple benchmarks. This comparison helps elucidate how scaling and different training resources influence the model's success.

This study also contributes to the BabyLM Challenge [2], aimed at training language models on developmentally realistic, small-scale datasets. The BoyBERTa model was developed and assessed using data provided by the challenge, serving as both a benchmark and a competition entry. This paper is organized as follows: the subsequent section offers an overview of the BabyLM Challenge and discusses prior research related to BabyBERTa's approaches to grammar learning. Following this, we detail the methodology employed in developing BoyBERTa, including the design and training processes. We then describe the experimental setup

*These authors contributed equally.

and present a comparative analysis of BoyBERTa against the baseline BabyBERTa and RoBERTa models, examining the impact of different scales and training resources on model performance. Finally, the discussion section considers the broader implications of our results and concludes with an overview of the main contributions and directions for future research.

Related Work

BabyLM Challenge

The BabyLM (Baby Language Modeling) Challenge is a research initiative aimed at fostering the development of NLP models trained on small-scale, child-directed corpora [2]. The challenge aligns with the goal of understanding how children learn language with limited exposure to data and aims to replicate this process computationally. The BabyLM Challenge provides curated datasets derived from child-directed speech, reflecting a realistic and developmentally appropriate language-learning environment. Participants are tasked with training and evaluating language models using this limited but high-quality dataset to simulate how children are exposed to language. The models are mainly evaluated on BLiMP (Benchmark for Linguistic Minimal Pairs) and SuperGLUE, which assess linguistic knowledge and general language understanding capabilities.

BabyBERTa

Huebner et al. [5] introduce BabyBERTa, a smaller version of the RoBERTa language model, in their study focusing on the grammatical knowledge acquired from language acquisition data. Despite the significant attention transformer-based language models have received in natural language processing (NLP), their potential for addressing key questions in language acquisition research has been largely overlooked. This study aims to bridge that gap by examining the grammatical knowledge BabyBERTa acquires when trained on a 5-million-word corpus that simulates the language input available to children between the ages of 1 and 6. The comparision between BabyBERTa and RoBERTa is shown as Table 1.

Table 1: Comparison between RoBERTa-base [7] and BabyBERTa [5]

	RoBERTa-base	**BabyBERTa**
Parameters	125M	5M
Data Size	160GB	0.02GB
Words in Data	30B	5M
Batch Size	8K	16
Max Sequence	512	128
Epochs	>40	10
Max Step	500	260
Hardware	1024x V100	1x GTX1080
Training Time	24 hours	2 hours
Accuracy	81.0%	80.5%

Using the behavioral probing paradigm, the authors discovered that BabyBERTa, which never predicts unmasked tokens, acquires grammatical knowledge comparable to that of the pre-trained RoBERTa-base model. Remarkably, BabyBERTa achieves this with approximately 15 times fewer parameters and using 6,000 times fewer words.

The findings have significant implications for building more efficient models and understanding the learnability of grammar from the limited input available to children. To further support research in this area, the authors have introduced a novel grammar test suite compatible with the small vocabulary of child-directed input, which they have made publicly available.

Curriculum learning

Curriculum learning is an approach to machine learning where the training examples are not presented randomly but organized in a meaningful sequence. This sequence starts with simpler concepts and gradually introduces more complex ones [1]. There are two proposed advantages to this approach:

1. Reduced training time: By not exposing the learner to noisy or difficult examples too early, it avoids wasting time on predictions that are likely to be inaccurate at the initial stages.
2. Better optimization: The curriculum helps guide the learner towards more favorable regions of the training space, avoiding local minima during the optimization process.

In summary, curriculum learning aims to enhance the efficiency and effectiveness of the training process by presenting examples in an organized, curriculum-like manner, starting from the simplest concepts and gradually increasing complexity.

Methodology

We are using BabyBERTa as our base model. We want to test whether adapting curriculum learning in this task will improve the model's capabilities for understanding language or not. We will train BabyBERTa on the strict-small track with a 10M word corpus under two scenarios:

Curriculum Learning: In this scenario, we will train the model using curriculum learning, where we start with easy tasks like child-directed speech, and then gradually move to more complex tasks like non-dialogue text. In particular, we train model with dataset with the following order: CHILDES, BNC, Switchboard, OpenSubtitles, Project Gutenberg, and Simple English Wikipedia. The meaning of each dataset can refer to Table 2.

Random Order: In this scenario, we will train the model without following a specific curriculum. The order of the corpus will be randomized, and the model will be trained on a mixture of easy and complex tasks simultaneously. Our random order resulted in the following: Switchboard, OpenSubtitles, Project Gutenberg, CHILDES, Simple English Wikipedia, and BNC.

After training, we will evaluate the model's performance on the BLiMP tasks (The Benchmark of Linguistic Minimal Pairs) using BabyLM evaluation pipeline 2023 [12] [4].

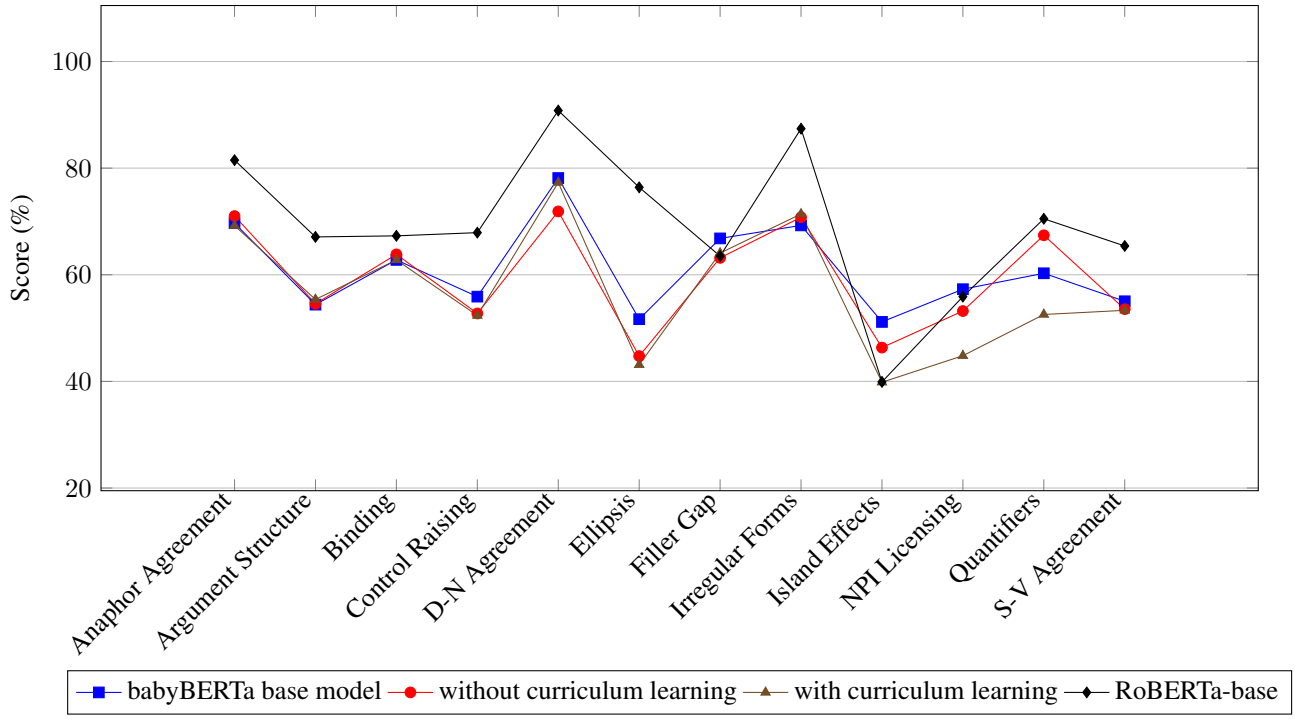

Figure 1: Model comparison on BLiMP tasks.

This evaluation will determine if the curriculum learning approach, where tasks are presented in a specific order, yields better results compared to the random order approach, where tasks are presented randomly. Additionally, we will compare the performance of our models against babyBERTa, which is our base model, and RoBERTa, which is the base model for babyBERTa.

Table 2: Data Source Weights and Domains of 10M word corpus

Source	Weight	Domain
BNC	8%	Dialogue
CHILDES	29%	Dialogue, Child-Directed
Project Gutenberg	26%	Fiction, Nonfiction
OpenSubtitles	20%	Dialogue, Scripted
Simple English Wikipedia	15%	Nonfiction
Switchboard	1%	Dialogue

Result and Discussion

Figure 1 showcasing the performance of four language models—babyBERTa base model, models trained with and without curriculum learning, and the RoBERTa-base model—across a variety of linguistic tasks offers a detailed look into the nuanced capabilities of each model. Each task represents a different aspect of language understanding, from syntactic structure to semantic processing, allowing us to pinpoint where each model excels or struggles.

Comparison of models

Across most tasks, the RoBERTa-base model stands out with consistently higher scores, suggesting a robust ability to handle complex linguistic challenges. This model particularly excels in Determiner-Noun Agreement, Irregular Forms, and Ellipsis, indicating strong grammatical and morphological understanding. The babyBERTa base model, while generally performing well, tends to lag behind the RoBERTa-base, especially noticeable in tasks requiring nuanced grammatical intuition such as Determiner-Noun Agreement and Ellipsis. The comparison between the models trained with and without curriculum learning reveals minimal differences, suggesting that the curriculum structure does not significantly alter performance in this dataset. Both curriculum models show similar trends, with slightly better performance in some tasks like Determiner-Noun Agreement and slightly lower in others like Control Raising, but these variations are not substantial enough to suggest a clear advantage of one training method over the other.

Does curriculum learning affect the performance?

The comparison of models trained with and without curriculum learning shows minimal performance differences across various linguistic tasks, as indicated by a line graph. Both models exhibit similar performance trajectories, with no significant variance suggesting a decisive impact from curriculum learning. For example, in Determiner-Noun Agreement, the curriculum model scores 77.3%, slightly higher than the non-curriculum model's 71.88%, but the margin is not substantial. In complex syntactic tasks like Control Raising,

the curriculum model scores 52.36%, slightly lower than the non-curriculum model's 52.74%. These observations imply that while curriculum learning may theoretically organize the learning process and aid in handling task complexity, its practical impact on performance is negligible in this dataset. This prompts further investigation into optimizing curriculum design and integration for more effective language model training.

How do scaling down the training data and using a developmentally plausible corpus impact the model's performance?

BabyBERTa, a scaled-down version of RoBERTa trained on a 5 million word corpus, aims to refine the model's learning focus with a smaller dataset. While BabyBERTa shows robust performance across several linguistic tasks, it generally falls short of the RoBERTa-base model's scores, which benefits from a larger, more diverse training set. This suggests that the reduced data size, while enhancing focused learning, limits exposure to diverse linguistic contexts, capping overall performance. In contrast, our models, based on BabyBERTa but trained on a 10 million word corpus, seek a middle ground by retaining focused learning benefits while introducing more linguistic diversity. Although the curriculum model mimics natural language learning sequences, practical outcomes show only marginal differences between curriculum and non-curriculum models, indicating limited advantages over the scaled-down BabyBERTa approach.

Conclusion

In summary, our analysis of the BabyBERTa model and its curriculum-based variations trained on differently scaled corpora reveals nuanced results. While the BabyBERTa, trained on a smaller 5 million word corpus, performs commendably, it does not reach the breadth of linguistic comprehension demonstrated by the RoBERTa-base model, which benefits from a much larger dataset. Meanwhile, scaling up the corpus size to 10 million words for models with and without curriculum learning shows only marginal improvements, suggesting that simply increasing data volume or incorporating a developmental training sequence like curriculum learning does not significantly enhance performance. This underscores the complexity of balancing data scale and training methodologies in language model development and highlights the need for precise optimization of these factors to truly benefit model capabilities.

References

[1] Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 41–48. New York, NY, USA: Association for Computing Machinery. ISBN 9781605585161.

[2] Choshen, L.; Cotterell, R.; Hu, M. Y.; Linzen, T.; Mueller, A.; Ross, C.; Warstadt, A.; Wilcox, E.; Williams, A.; and Zhuang, C. 2024. [Call for Papers] The 2nd BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. arXiv:2404.06214.

[3] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

[4] Gao, L.; Tow, J.; Biderman, S.; Black, S.; DiPofi, A.; Foster, C.; Golding, L.; Hsu, J.; McDonell, K.; Muennighoff, N.; Phang, J.; Reynolds, L.; Tang, E.; Thite, A.; Wang, B.; Wang, K.; and Zou, A. 2021. A framework for few-shot language model evaluation.

[5] Huebner, P. A.; Sulem, E.; Cynthia, F.; and Roth, D. 2021. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In Bisazza, A.; and Abend, O., eds., *Proceedings of the 25th Conference on Computational Natural Language Learning*, 624–646. Online: Association for Computational Linguistics.

[6] Jawahar, G.; Sagot, B.; and Seddah, D. 2019. What Does BERT Learn about the Structure of Language? In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. Florence, Italy: Association for Computational Linguistics.

[7] Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

[8] Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683.

[9] Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT Rediscovers the Classical NLP Pipeline. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601. Florence, Italy: Association for Computational Linguistics.

[10] Tu, L.; Lalwani, G.; Gella, S.; and He, H. 2020. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. arXiv:2007.06778.

[11] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[12] Warstadt, A.; Choshen, L.; Mueller, A.; Williams, A.; Wilcox, E.; and Zhuang, C. 2023. Call for Papers – The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. *Computing Research Repository*, arXiv:2301.11796.