Investigating the Impact of Pre-training on Child Language Data for BERT Models and its Performance on Downstream Grammatical Tasks

Shiven Taneja and Priyanka Raj

Abstract

The transformer architecture, specifically BERT (Bidirectional Encoder Representations from Transformers), has recently become a popular choice for Natural Language Processing (NLP) tasks. This is due to its ability to capture long-term dependencies as well as contextual information. This paper aims to investigate the influence of pre-training datasets on BERT models, focusing on child language data, which is characterized by variability and variations from common linguistic structures. By replicating a BERT model, we were able to examine the loss difference between different pre-training datasets. We then evaluated the performance of a Hugging Face Bert model pretrained on these different datasets (both child oriented and adult oriented) on a Masked Language Modeling task with different grammatical sentence structures. We found that all the models were able to capture the correct grammatical parts of speech but failed to consider specific contextual information which is likely due to insufficient training time. The large differences in the sizes of the datasets may have impeded our ability to isolate the effects of training on child-directed inputs and thus further investigation is necessary.

Keywords: Machine Learning; Transformers; Natural Language Processing; Large Language Models; Artificial Intelligence; Masked Language Modelling; BERT; BabyBERTA

Introduction

In the field of Natural Language Processing (NLP), several models have been developed in an attempt to capture language and perform relevant human-like tasks. What started with the usage of feedforward networks and a recurrent neural network's (RNN) ability to use positionality in past input sequences to predict that of future output sequences, has progressed to incorporating attentional mechanisms into this architecture. Such that RNNs would be able to focus on certain parts of an input sequence when predicting parts of the output sequence.

Alternatively, Vaswani et al. (2017) proposed a Transformer architecture, which relies solely on self-attentional mechanisms. That being that the model can attend to different parts of an input sequence to create a representation for each component in the sequence. Conventionally, transformer models have two essential components: an encoder and a decoder. The encoder creates a sequence of hidden states to capture the meaning of an input sentence. The decoder then uses these hidden states to generate an output sequence.

Unlike previous neural networks, the self-attentional mechanisms of transformer architecture allow the model to capture long-term dependencies between parts of a sequence to make it useful for downstream NLP tasks, such as text summarization or sentence completion.

BERT (Bidirectional Encoder Representations from Transformers) is a type of transformer-based neural network for NLP tasks. Proposed by Devlin et al. (2018), it has become one of the most popular and widely used NLP models.

Unlike traditional transformer architectures, BERT consists of solely encoder layers. Therefore, there is no decoder that is trained to predict the next word in a sequence, but rather a BERT model is pre-trained on a large corpus of text through two unsupervised language learning strategies: masked language modeling (MLM) and next sentence prediction (NSP).

During MLM, roughly 15% of tokens from the input text are randomly masked and the model is trained to predict these missing tokens based on the other tokens. Thus, representing the architecture's bidirectional nature and parallelization, which affords the learning of contextual information and a capturing of deeper semantic relationships between words.

During NSP, the model is trained to predict whether a pair of sentences are consecutive, thus allowing the model to learn the relationship between sentences and global contextual coherence of a text.

The importance of this pre-training process cannot go overlooked, since its essential for the success of these models to capture and use linguistic representations for downstream NLP tasks. Evidently, the characteristics of the data with which the model is pre-trained on is critical, since it influences the linguistic representations that the model learns and how well it performs on downstream tasks.

A recent area of research is the impact of pre-training models on child data. Child language is considered highly variable, as it includes grammatical errors and deviations from the common linguistic structures found in everyday text. Therefore, we hypothesize that pre-training models on child data may improve the robustness of the model, its ability to handle variability in input text, as well as diversify the grammar learned. Exploring the influence of pre-training data can improve the performance of NLP models such as BERT on different downstream applications.

Thus, this paper aims to first replicate a BERT model, to ensure that the different datasets are pre-processed in the same manner, and to investigate loss differences between different pre-training datasets. Furthermore, it also aims to explore how different pre-training datasets may influence the downstream task of sentence completion, based on

grammatical sentence structures relevant to children's language capabilities.

Related Work

The BERT model was introduced by Devlin et al. (2018) and several studies have worked on extending and improving the performance of the BERT model by manipulating different aspects of the model, including pre-training dataset characteristics.

Related to the size of the data the model is pre-trained on, it's common to assume that the more data a model is pre-trained on, the better the performance on downstream tasks. However, studies have questioned whether this is always the case. Sanchez and Zhang (2022) explored the effects of within-domain corpus size on pre-training BERT models. They conducted a series of experiments where they pre-trained their BERT model with different sizes of biomedical corpora. Their results showed that pre-training on a relatively small amount of within-domain data, even with limited training steps, outperformed on downstream domain-specific tasks compared to models that are commonly pre-trained on general corpora and then merely fine-tuned for the specific task. Therefore, specificity and quality of the pre-training data may be more relevant than the quantity of data being used.

Beyond exploring the effect of corpora size, we can therefore look at the qualitative aspects of the data. To propose a more efficient and effective pre-training approach for BERT models on NLP tasks, Huebner et al. (2021) trained their BERT model on a corpus of language acquisition data to represent the input available to children between the ages of 1 to 6. Their results found that the model pre-trained on the language acquisition data was able to attain grammatical knowledge at the level of standard BERT-base models and did so with fewer parameters and much less data. Thus, similarly to the previous effect of corpus size, this study calls for more efficient models as well as learnability of grammar from language acquisition input usually available to children.

Therefore, we aim to focus on the domain-specific pre-training of our BERT model on a relatively limited collection of child-directed data and how this specific pre-training approach may influence the performance on grammatical downstream tasks.

BERT Model

The first portion of this study aims to build a BERT model and pre-process a collection of child data and adult Wikitext data such that they can be compatible with the model. Loss function visualizations are then generated for both MLM and NSP performance to evaluate the built model (https://github.com/shiven-taneja/BERT_0.1).

Dataset Pre-Processing

The first stage before building the model is preparing the dataset for pre-training. This section involved splitting the dataset on sentences, tokenizing the data to create a vocabulary of word-token pairs and creating a training dataset. In order to do so, we create our pre-processed dataset by extending Pytorch's Dataset class.

We chose to compare how child-directed data and standard Wikitext data would compare on the MLM and NSP losses of our model. Our child-directed data comprised of the CHILDES dataset, a dataset of excerpts from children's books (provided by Hugging Face) and a dataset of children's stories (provided by Hugging Face). All of which were collected and processed altogether in our dataset_childes.py. Our standard data was Wikitext-103-v1 (provided by Hugging Face) and was pre-processed in our dataset_wikitext.py file. The details of both the Child-Text and Wiki-text is provided in Table 1.

Each of the dataset_childes.py and the dataset_wikitext.py split the dataset based on sentences, and using this creates the vocabulary of the model for word-token pairs. They then also create the training dataset, which entails pre-training on MLM and NSP tasks. For MLM training, we randomly mask 15% of tokens (i.e. replace them with [MASK]) and use unmasked sentences to train the model for MLM tasks. As for NSP, the model takes a complete excerpt of more than one linguistic sentence to mark whether a sentence follows another or whether it's a random sentence. This is to train on predicting the order of sentences.

Build BERT Model and Trainer

The BERT model we built has 3 embedding layers which transform the words into vector representations. These 3 layers are the Token Embedding Layer, which encodes the word tokens, the Segment Embedding Layer, which encodes whether it is part of the first or second sentence, and the Positional Embedding Layer which encodes the position of the word in the sentence.

The model also has a Multi Head Attention Layer which is multiple Attention Heads working in parallel. These Attention Heads helps the model understand and process contextual information in sentences by weighting the importance of each word in relation to others in the sequence, creating a contextual representation of the input. By having multiple Attention Heads, the model can focus on different aspects of the data simultaneously. Each model consisted of 6 attention heads through which the data is then given to a feed-forward neural network. This takes place in the Encoder Layer and our model consists of 4 identical Encoder Layers. All of these layers are then combined in order to produce the output for the MLM and NSP pre-training tasks.

We also had to create a BertTrainer Class which initializes and pretrains the model on the dataset. In order to train the model, we used Negative Log Likelihood for the MLM task, Sigmoid Binary Cross Entropy Loss for the NSP task and an Adam optimizer. The models' parameters are then updated using the loss calculated. We trained our models for 4 epochs with a batch size of 12, an embedder size of 64 and a hidden size of 36.

NSP and MLM Loss

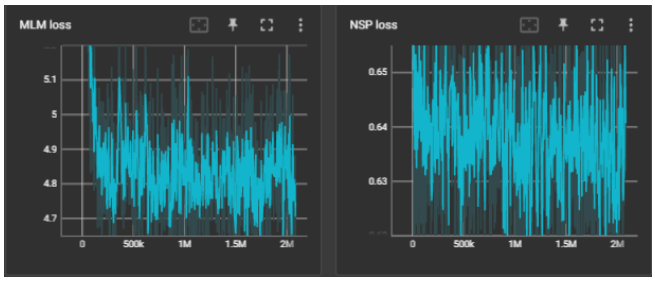

Figure 1: Wikitext-103 Loss Functions

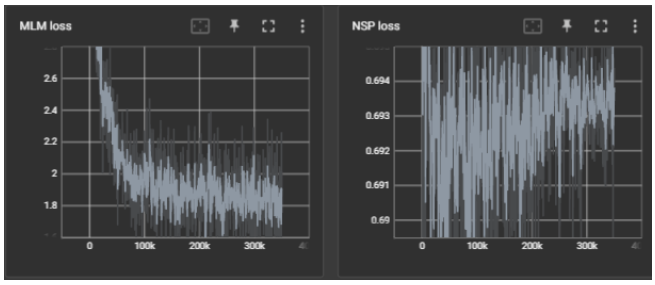

Figure 2: Child-text Loss Functions

From Figure 1 and Figure 2, we can see that the MLM loss for both the Wikitext-103 dataset and the Child-text dataset converges over time. Despite this, we can see that the Child-text converges a lot faster than the Wikitext-103 dataset despite being trained for the same number of epochs (4). The better convergence of the Child-text MLM loss shows that the model is learning more effectively. The difference between the MLM loss convergence of the models is likely due to the difference in size of the datasets as the Wikitext-103 dataset is over 5.5 times bigger than Child-text (Table 1).

Surprisingly, we see a different result for the NSP loss. For both models, it seems that the NSP loss diverges. This is seen even more in the Child-text Loss. The NSP loss is supposed to converge as this shows improved performance, but we see the opposite result here which may be due to insufficient training time or an unsatisfactory learning rate.

Hugging Face BERT Model Performance

In order to effectively test and evaluate the performance of the BERT models, we utilized Hugging Face's (HF) 'BertForMaskedLM' and pre trained it on the different datasets. This would allow us to isolate the effects of the different datasets on a MLM task.

Model Evaluation and Analysis

We used a grammar test we found online that tests each model's performance on 3 questions from 5 different categories taken from popular online grammar tests (Sangani, 2021). These 5 categories are Prepositions, Articles, Question Tags, Opposites, and Proverbs. We trained 3 different models on different datasets and used the 'bert-base-uncased' model as a comparison. The comparison of these models can be shown in Table 1.

Table 1: A comparison of all HF models

	Child-Text	**CHILDES**	**Wiki-text**
Data Size	95.9 MB	21.5 MB	529.6 MB
Words	16118638	4327167	86537576
Distinct Words	235068	25380	226110
Epochs	4	4	4
Training Time	3.8 hours	2.3 hours	7.7 hours

As we can see, the Wiki-Text dataset is over 24.5 times larger than the Childes dataset and over 5.5 times larger than the Child-Text dataset.

Grammar Test

In order to test the language capabilities of the models trained, we used a MLM task of 5 categories. The questions for each category are:

Prepositions
- 'Using your cell phone while driving is [against] the law.'
- '[Although] she's a little shy, she's a wonderful person once you get to know her.'
- 'We drove [along] the coastline of California.'

Articles
- 'Carl lives alone in [a] one-bedroom apartment.'
- 'The test results will be available in about [an] hour.'
- 'He is [the] best of what we have got.'

Question Tags
- 'You would'nt like to invite my dad, [would] you?'
- 'You do go to school, [don't] you?'
- 'They will wash the car, [wont] they?'

Opposites
- 'Black is the opposite of [white].'
- 'Love is the opposite of [hate].'
- 'An atheist is the opposite of a [Christian/Muslim/Hindu etc…].'

Proverbs
- 'Better late than [never].'
- 'Slow and [steady] wins the race.'
- 'When there is a [will] there is a way.'

In this MLM test, the words in square brackets were masked and inputted as [MASK]. The words in the brackets are the ground truth values we created and through this we compared the results of the model to determine their performance.

Not only did we evaluate whether the answer was correct, we also analyzed the model's 'confidence' (score) in its answer which is given as a ratio from 0 (no confidence) to 1 (high confidence). The more confident the model is of its answer, the higher this score will be.

Model Evaluation

In order to evaluate the performance of all the models, we will first discuss how they fared on the MLM tasks overall.

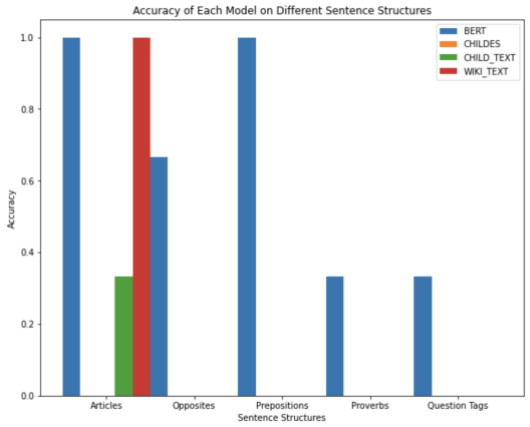

Figure 3: Accuracy of Each Model on Different Sentence Structures

As we can see from Figure 3, the BERT model performed exceptionally well for the Articles and Prepositions sentence structures and performed reasonably well in all the other categories. On the other hand, the Wiki-Text model performed equally well on the Articles category but was unable to correctly predict the masked word for any of the other category. The Child-Text model performed reasonably well for the Articles sentence structure but incorrectly predicted the masked words for every other category. The CHILDES model performed poorly across every category and did not predict the correct masked word for any of the 15 sentences.

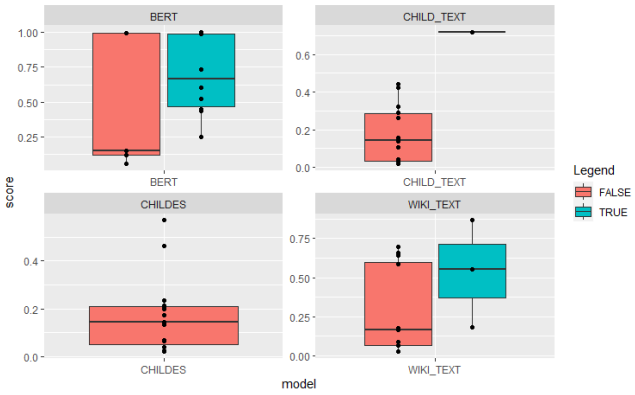

Figure 4: Boxplots showing the confidence scores of answers that were both True and False

Figure 4 analyses the confidence scores of the models answers that were both True and False. Through this we can not only analyze the performance of the model, but we can also look at how confident each model is of its answers. Looking at the BERT model, we can see that the mean confidence score for False answers was much lower, but for one of its answers that was incorrect it still gave an extremely high confidence score. This led to the interquartile range (IQR) of the False answers to be extremely wide. For the True answers, we can see the mean is a lot higher meaning that for answers it had a high confidence in, they were more likely to be right. For the Child-Text data on the other hand, we can see a much smaller range in values for its False answers and its highest confidence was for the only question it correctly unmasked. The CHILDES model had no correct answer and for most of its wrong answers, it had a low confidence score showing that it struggled to predict the masked word for the sentences provided. The Wiki-Text model had a large spread of confidence scores for both False and True answers. Despite this, the mean score for False answers was over half as small but there were still 4 False answers that had a confidence score greater than the mean True confidence.

Sentence Structure Evaluation

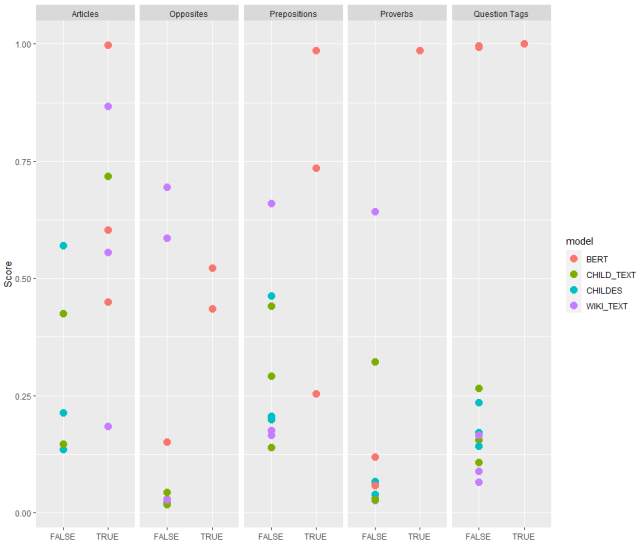

Figure 5: Confidence scores based on true or false answers, across the different sentence structures

From Figure 5, we can notice again the greater accuracy of the original BERT dataset given the number of orange points on the True data line for each sentence structure. Similarly, we see that most of the other datasets, especially CHILDES, have lower accuracy by having most points in the False data line, except for some exceptions in the structures involving articles.

Interestingly from Figure 5, we can also see how the confidence levels varied for each model across different sentence structures.

One pattern to notice is the high levels of confidence for wrong answers. For the sentence structures involving articles, opposites, prepositions and proverbs, some models have a confidence level of above 50% in thinking that a wrong answer was right. Especially for articles and prepositions where most of the models that got an answer wrong were

relatively sure it was right. Interestingly even the original BERT dataset exhibited this pattern in the sentences involving question tags and with the highest amount of certainty.

Another pattern to consider is the spread across confidence levels within a sentence structure. For example, for sentences with articles we see that even for both True and False answers, the models had varied confidence levels that spanned the spectrum. However, we see that for sentences involving question tags, most of the models converge on either having very low confidence or very high confidence.

Discussion

A qualitative evaluation of the actual outputs can help explain such patterns and perhaps give light as to the structures captured by particular datasets.

As for the pattern of having higher confidence levels for wrong answers (i.e. the False data line), we can take a closer look at certain examples from the outputs in the output.txt file. Particularly we notice that for prepositions, even though a model's answer did not match the ground truth, it did have the correct part of speech (POS). For example, for the sentence "Using your cell phone while driving is [against] the law", all models, except for the original BERT data, put "in" instead of against. Where this makes grammatical sense in terms of its POS, it fails to capture the domain specific concept that using your cell phone while driving is illegal. This is even the case for the Wikitext data, which is not child-based.

Similarly, if we look at another preposition example, "[Although] she's a little shy, she's a wonderful person once you get to know her", we see that the three models that got it wrong input "and", "because", and an open quotation mark. Especially in the case of a "because" input, the POS seems to logically follow but seems to not consider the entirety of the sentence or the complex sentence structure.

The third example in the preposition sentences, where it's "we drove [along] the coastline of California", the three models (except for the original BERT data) input "all" and "to", which both grammatically (in terms of POS) and (for the case of "to") contextually makes sense.

Overall, we see that the pattern of having higher confidence levels despite choosing the wrong answer reflects correct grammatical POS capture. However, it also brought to light flaws of the Child-text, CHILDES and Wikitext models in considering specific domain-contextual information and recognizing certain sentence structures.

As for the spread of confidence levels across models on certain sentence structures compared to others, we can look at the sentences involving articles and those involving question tags.

One of the articles sentences is "Carl lives alone in [a] one – bedroom apartment", which was responded correctly by the Wikitext and original BERT models, but the CHILDES and Child-text models responded with "the". If there was reference to a specific one-bedroom in this context, then this input would be correct. Thus, an example of correct grammatical POS capture, but lack of consideration of grammatical context (i.e., whether a referred subject was mentioned previously, thus to suggest using a definite article like "the"). Nonetheless, the grammatical correctness and interchangeability of certain articles like "a" or "the" based on grammatical context could likely be the reason behind models distributing certainty relatively equally across different articles.

Conclusion

Initially, building and pre-training a limited BERT model from scratch gave promising results to be able to converge on MLM loss. Further exploration using a standard and robust BERT model (by Hugging Face) to evaluate the influence of different pre-training datasets was then able to give insights as to what was captured and missed by the pre-training datasets (given that the based model was kept the same). Overall, integrating both the qualitative and quantitative evaluations of each model's outputs and the pattern of overall outputs, we can identify that all models were able to recognize correct grammatical POS, which explains patterns of high levels of certainty and fluctuations in certainty. However, the qualitative evaluations, in particular, gave light to the lack of contextual consideration both in terms of semantic context and grammatical context.

The reasoning behind these conclusions can come from certain limitations in our implementation. one, the training time was extremely limited which may have led to underfitting and incomplete learning. The models may not have had enough time to optimize its weights and learn the complex relationships in the data. This may have led to the suboptimal performance on the downstream task that we saw. Furthermore, the small size of the datasets may have impacted the model's ability to generalize to new data as it may not have been exposed to enough examples during pre-training to adapt to different scenarios.

To take the research further, it would be beneficial to properly isolate the effects of the quality of the data such that size is not variable. This would allow us to root the issues of the downstream tasks. Furthermore, in terms of the downstream MLM task, we could be able to incorporate multiple sentences such as to explore and isolate the performance of the model on grammatical contexts (e.g., using "a" vs "the").

In terms of exploring related concepts, it would be interesting to investigate the ability for models to capture common child disfluencies, such as code-switching between terms of different languages. Similarly, whether such patterns of language acquisition differ across cultures and languages. Additionally, there could be influences of language acquisition abilities on other cognitive processes or downstream tasks such as question-answering or reasoning.

References

Bert-base-uncased · hugging face. bert-base-uncased · Hugging Face. (n.d.). Retrieved April 11, 2023, from https://huggingface.co/bert-base-uncased

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*. https://doi.org/10.18653/v1/n19-1423

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). Babyberta: Learning more grammar with small-scale child-directed language. *Proceedings of the 25th Conference on Computational Natural Language Learning*. https://doi.org/10.18653/v1/2021.conll-1.49

Sangani, R. (2021, September 9). How do transformers fare on a grammar test? Geek Culture. https://medium.com/geekculture/how-do-transformers-fare-on-a-grammar-test-d3b7b26e5004

Sanchez, C., & Zhang, Z. (2022). The Effects of In-Domain Corpus Size on pre-training BERT. https://doi.org/https://doi.org/10.48550/arXiv.2212.07914

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762