BERT and Child Language Acquisition

New York University Center for Data Science
DS-GA 1016 Computational Cognitive Modeling
Wendy Hou (wh916@nyu.edu)
Gabriella Hurtado (gh1408@nyu.edu)
Stephen Roy (sr5388@nyu.edu)

Abstract

BERT, short for the Bidirectional Encoder Representations from Transformers, is a new and unique language model because of its ability to learn unsupervised and bidirectionally, thus performing outstandingly in various language understanding benchmarks. This study examines the similarities and differences between BERT and child language acquisition by testing BERT on Kumon reading tasks that are designed to evaluate and teach children of different age groups basic English grammar. Compared to Kumon students' performances, BERT generally performs worse given the same task. The three reasons highlighted in this paper for BERT's underperformance are: 1. BERT fails to overcome its prior bias learned during pre-training and favors previously common candidates instead of the correct predictions for the specific Kumon tasks; 2. BERT analyzes each masked word independently and fails to relate them; 3. BERT performs poorly when asked to predict phrases (adjacent masked words).

Keywords: BERT; deep learning; neural networks; child language acquisition, constraint learning

Introduction

BERT is the latest state-of-the-art language model developed by Google in 2018. What is ground-breaking about BERT is that this is a model that learns by simultaneously conditioning on the elements to both the left and right sides of the target (Devlin et al., 2019). Unlike previously unidirectional language models, BERT is able to obtain more comprehensive information and richer context while learning, thus generally performs better than its predecessors, like Long short-term memory (LSTM) neural networks.

LSTMs are recurrent neural networks that have the ability to learn a significant amount of grammatical structure, like subject-verb agreement. In strongly supervised settings, LSTMs can achieve less than a 1% error rate for various tasks even though the model does not have built-in hierarchical representations. However, LSTMs error rate increases when the difficulty and complexity of the sentences increases (Linzen et al., 2016). Based on the studies on LSTMs, later studies replicate similar language tasks using BERT to study how BERT learns the structure of language.

For example, each layer's performances in different language tasks are different. Features of language, like subject-verb agreement, are captured well by the middle layers, while the deeper layers perform better at semantic tasks, and the shallower ones perform better at surface information like word count (Jawahar et al., 2019). This has a general resemblance in child language acquisition, where children first learn surface information about languages, like sounds, syntax and grammar rules like word orders next, and semantics later (Professor

Ailis Cournane, personal communication, May 1, 2020). However, more evidence is needed to argue that the different layers in BERT have connections to the different age groups or stages in childhood language acquisition. Although retraining BERT and visualizing each layer's performance is not tangible for this study, it does tackle the problem of how BERT performs on language structure understanding in order to compare BERT's learning with how children learn language structures. In order to achieve this, this paper evaluates BERT's performances on a set of Kumon reading tasks that are designed to evaluate and teach children of various age groups about basic English grammar.

Child language acquisition is an important field in cognitive psychology. In the specific field of grammar acquisition, the theoretical approaches can be categorized into two groups: the generativist approaches and the constructivist approaches (Ambridge et al., 2011). The generative approach believes children learn grammar as general rules that can be summarized using a phrase-structure tree (Ambridge et al., 2011). Various studies have shown that BERT encodes in an approximation of the syntax tree which is the core of the generativist approaches (Chrupała et al., 2019; Coenen et al., 2019; Hewitt et al., 2019; Jawahar et al., 2019). In BERT, the tree distance roughly corresponds to the square of Euclidean distance (Hewitt et al., 2019) based on Pythagorean embedding (Coenen et al., 2019).

On the other hand, the constructivist approaches suggest that the end goal of child grammar acquisition, adult fluency in the language, is a set of constructions or patterns that are closely associated with the meaning of sentences (Ambridge et al., 2011). By noticing the correlation between a specific pattern and the meaning of it, children then start to learn about the constructions. Interestingly, these constructions have different levels of abstraction and, by practicing similarly structured sentences with variations, children eventually learn the most abstract level of construction which is grammar (Ambridge et al., 2011).

At the same time, while BERT can be trained and retrained on incredibly massive text corpora (e.g. Wikipedia and GLUE), few studies have trained it on collections of purposefully similar sentences with small alterations, which is an extremely common narrative style in children's literature. Therefore, BERT differs from the constructivist approaches since the essence of learning the constructions is through repeating sentences of similar structures with small variations. Nonetheless, the similarities between BERT and the constructivists is that the constructivist approaches specifically address the importance of semantics in child grammar acquisition (Ambridge et al., 2011). As the first successful bidirectional language model, BERT clearly shares this characteristic by considering as much context as possible and constructing subspaces to accurately capture related semantic information (Coenen et al., 2019; McClellanda et al., 2019).

Despite these similarities between BERT and theoretical approaches towards child grammar acquisition, there are some crucial differences that should not be overlooked. First, as mentioned previously, BERT is trained on massive text bodies like Wikipedia. However, these are not the typical literature presented to children in everyday life. Additionally, even though children growing in multilingual environments are not rare, BERT is trained on more than 70 languages; this

degree of language training is extremely rare for people.

More importantly, human brains are much more complicated and flexible than any language model available and that people don't only stick to one model when learning languages. As an example, humans have countless complementary learning systems that are simply not available in BERT. Like many other language models, BERT cannot access the specifics of the prior information when a text is replaced by a new one, while humans can learn from information presented just once in the past (McClellanda et al., 2019). Moreover, people do not only rely on text to learn languages; senses like vision, hearing, gestures all play essential roles in child language acquisition, but they are not available for BERT (McClellanda et al., 2019).

Additionally, it is important to keep in mind that the generativist and constructivist views are largely based on the natural learning environment of children's language acquisition (Professor Ailis Cournane, personal communication, May 1, 2020). Although the Kumon dataset used in this experiment is related, it represents constraint learning in children and does not directly reflect a natural learning environment for children learning languages.

Methods/Models

Data. The tasks used to test BERT are derived from Kumon reading workbooks. Workbook levels AII, BI, BII, CI, and CII were consulted, the difficulty levels of which correspond to first- to third-grade reading comprehension. According to Kumon, workbook level AII teaches students to "recognize a sequence of thoughts developed within a short paragraph," levels BI and BII focus on students' abilities to "identify subject and predicate in longer sentences...define words using context clues...and to compare and contrast actions," and levels CI and CII have students develop the skills to write "complete sentences independently," and "identify subjects, verbs, and objects," (Kumon Institute of Education, 2010).

Each Kumon workbook level is composed of twenty groups of ten worksheets each (200 pages total), and students who are enrolled in Kumon are assigned between five and ten pages of homework to complete each day. For example, a Kumon assignment may be to complete pages AII 135 to AII 140, or CI 20 to CI 30. Depending on its difficulty, a singular worksheet may contain anywhere from two to thirty exercises, although the majority of the worksheets referenced in this study were made up of just four tasks.

Performance on Kumon classwork is graded on a custom scale. A student receives a possible grade of 100%, 90%, 80%, 70%, or 69% for their efforts on a worksheet. These percentages are not calculated directly from the proportion of right to wrong answers on a worksheet, but rather predetermined by Kumon. Worksheets with four tasks, for example, are graded as follows: 100% for no wrong answers, 80% for one wrong answer, 70% for two wrong answers, and 69% for three or more wrong answers. It is important to keep in mind that the lowest score a student may receive on a worksheet is 69%.

Common exercises in the AII and BI workbooks are timed reading tasks, spelling and tracing tasks, vocabulary matching exercises, and fill-in-the-blank questions. In the BII, CI, and CII workbooks, True/False questions, word tense modification, pronoun identification, and fill-in-the-blank questions are most prevalent. However, the tasks in this paper are limited to fill-in-the-blank questions

with no pictorial components to best replicate the masked language modeling BERT performs.

Tasks were chosen based on their suitability and average student performance. Suitable tasks included those that could be readily represented as masked language modeling exercises and that belonged to a worksheet composed only of similarly suitable tasks. This second condition was particularly important to this experiment, as it ensured that fair comparison between BERT and human performance was possible.

To best capture this comparison, the study elected tasks that best spanned the range of average student performance. Among the exercises chosen, the highest average student score was 96.9%, and the lowest was 72.8%. Levels AII and BI had the highest mean scores, with averages of 95.4% and 91.9% respectively. Levels BII, BI, and CI had similar average scores, with the lowest being level CI at 88.5%.

Kumon levels BI and BII, which cover second-grade comprehension, contained the greatest proportion of usable tasks, as well as the greatest variability in student performance among these tasks. Across BI and BII tasks only, the highest average student score was 93.3%, and the lowest was 77.6%. The average age of students enrolled in levels BI and BII are 8.5 and 8.7 years of age respectively.

Model and Methods. The huggingface transformers python module was used to implement BERT using the bert-base-uncased model with default (pretrained) model weights. To test BERT, the Kumon dataset was first transformed using the BertTokenizer, before applying both the BertModel and BertForMaskedLM (masked language modeling) models. Doing so provided easy access to the standard BERT output layer as well as mask predictions. Although there is no basis for comparison to the Kumon students, investigating BERT's layer attention could provide insights into explaining the model results with further analysis.

The bulk of the assessment thus relied on testing BERT's predictions on context, task pairings and comparing BERT's results to both the actual values and the average Kumon student performance. This was accomplished using sorted output from BertForMaskedLM and applying the Kumon grading function. Complete code, cleaned data, and a demo notebook is available for review on github.

Results

General EDA on Kumon students' and BERT's scores indicates that BERT routinely underperformed than the Kumon students. The only exception was on AII 148, where Kumon students' average score was 90.5% while BERT's score was 100% (Figure 1).

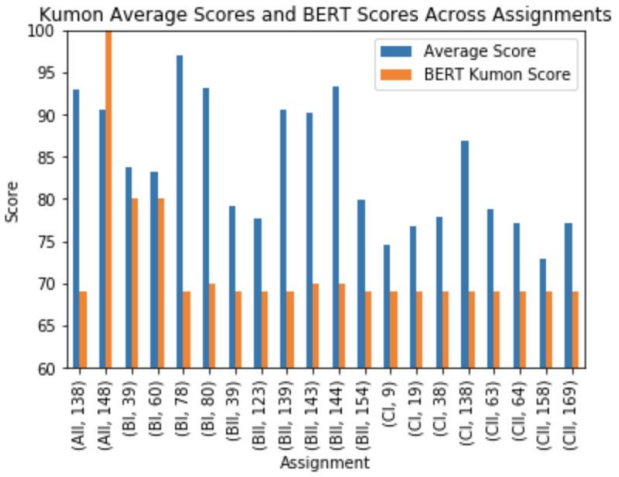

Figure 1. Comparison between Kumon students' average scores and BERT's score for each Kumon worksheet.

Otherwise, BERT's performance rarely exceeded 80% and constantly scored 69% on each worksheets when the average scores of

Kumon students ranged between 72.8% and 96.9%.

The mistakes made by BERT can be categorized into three general types: 1. BERT failed to overcome its prior bias learned during pre-training and leans towards commonly seen predictions instead of the correct predictions for the specific Kumon tasks; 2. BERT analyzed each masked word independently and failed to relate them; 3. BERT performed poorly when asked to predict phrases or sentences (adjacent masked words).

For the first type of mistake where BERT falsely considered its prior bias for predictions. A great example is from BI level: *"[CLS] When the sun started setting, she picked up her shoes, found her backpack and caught the next spacebus home. 'Where are you?' asked her mother. 'Are you all right? Your school called and I was so worried!' 'I'm fine, Mom,' Astrid said, handing her a bunch of daisies. 'I've been thinking… could we move out of the city? I know of a great place.' [SEP] What did Astrid hand her mother? Astrid handed her mother a [MASK] of [MASK]. [SEP]"* and the answers were ['bunch', 'daisies'] while BERT predicted ['bunch', 'flowers']. After examining the full list of possible predictions provided by BERT, 'daisies' was found at the 16th position, after 'tissue', 'roses', 'chocolate' and etc. This is possibly due to the fact that these other predictions are common expressions following the phrase "a bunch of", although they may not be the correct answer for this particular task. While the Kumon students' average score for this particular question is unknown, the students' average score for this worksheet is 93.2%, while BERT's score is 70%. It shows that children, while learning languages, have an easier time tuning into the specific task that is assigned to them. That is, while BERT evaluates possible answers based on its prior knowledge during training, humans understand that regardless of the frequency of a specific phrase or combination of words they have seen in their learning process, the answer needs to make sense for the specific question in order to be correct.

An example for the second type of mistake is the following BII task: *"[CLS] Leaf Tail shot from the bower, and scrambled away as fast as he could. Finally he stopped to rest on an old log. But as he lay there his toes disappeared. And his tail was gone, too. The motley log was like his own skin. [SEP] What two things disappeared as Leaf Tail rested on the log? His [MASK] disappeared and so did his [MASK]. [SEP]"*. The correct answers were ['toes', 'tail'], and BERT predicted ['tail', 'tail']. While 'tail' was one of the right answers, the question asked what two things disappeared and BERT failed to realize that 'tail' should not be the answer for both masked words (Figure 2). Judging from the 90.5% average score for this question, the Kumon students performed relatively well on this question while BERT's score is 69%. This indicates a possible discontinuity between masked words for BERT since it provided two almost identical lists of predictions when the questions clearly asked for two different answers. To humans, however, the key word "two" in the question will trigger a response that if one thing is chosen for the first blank, it cannot be a candidate for the second one, which contributed to a higher score among the Kumon students than BERT.

```
[[['tail' '11.373841']  [['tail' '11.140414']
  ['feet' '10.427265']   ['feet' '10.839702']
  ['toes' '9.4515915']   ['toes' '9.623857']
  ['head' '9.419552']    ['ears' '9.380962']
  ['legs' '9.007271']]   ['legs' '9.165258']]]
```

Figure 2. BERT's top 5 predictions for the two masked words in this example.

The last type of mistake is that BERT performs poorly at predicting phrases and sentences. The CI level tasks are mainly predicting phrases and sentences and an example is, *"[CLS] I pressed the emergency stop button! My mother looked shocked, but I pointed and she understood. The injured dog turned toward the subway and the headlights shone into its sad eyes. My mother made another announcement. 'We'll be delayed for a few moments. There's a dog on the tracks!' A gasp was heard throughout the cars. [SEP] What did Jackie's mother make? She [MASK] [MASK] [MASK]. [SEP]"*. The correct answer should be ['made', 'another', 'announcement'] which formed a meaningful phrase. However, BERT's predictions were ['was', 'was', 'again'] which were neither grammatically correct nor meaningful. It seems like CI is a level that both the Kumon students and BERT struggled with, both scoring between 70% and 80%. Out of the eight questions presented in CI, BERT only answered one question correctly, and its predictions made the least sense compared to all other levels, including the harder CII level. BERT appeared to underperform in phrase/sentence predicting, possibly because, during pre-training, the masked words in sequences were chosen randomly and independently instead of in a pattern. This limited BERT's chance to learn how to predict phrases or sentences that need to follow a set of grammatical rules to make sense. BERT's performance for this specific kind of task may be improved by using a different class implementation such as BertForNextSentencePrediction instead of BertForMaskedLM, which is the baseline model used in this experiment.

Discussion

The paper identifies three main mistakes BERT makes in task-oriented language learning, which reflects constraint learning in child language acquisition. Despite the various similarities discussed in the introduction between BERT and child language acquisition, BERT's performances on the Kumon tasks presented in this study is worse than expected. It is therefore possible that, while BERT does reflect certain key features found in child language acquisition in a natural learning setting, when not retrained on the particular dataset, BERT does not perform as well as children when it comes to constraint language learning, like taking exams. This is, again, possibly due to the complexity and diversity when it comes to the signals and environments children have access to when learning languages that are simply not available to BERT.

A specific limitation of this experiment is that only the pre-trained BERT model was introduced. This default model is trained using text that is written and consumed by adults; in contrast, the Kumon tasks used to test the model were developed particularly for child language acquisition. By retraining BERT using the Kumon reading material, or similarly focused texts, there might be a more fair comparison between Kumon students and BERT task performance, thus providing clearer insight on how the acquisition of language differs between the two. Further analysis can also be done by looking at larger sets, expanding modeling to include next sentence prediction, and deeper attention analysis on the tasks BERT struggled with most.

References

Ambridge, B., & Lieven, E. V. M. (2011). Child Language Acquisition: Contrasting

Theoretical Approaches. Cambridge University Press.
Chrupała, G., & Alishahi, A. (2019). Correlating neural and symbolic representations of language. *arXiv preprint arXiv: 1905.06401*
Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F., & Wattenberg, M. (2019). Visualizing and Measuring the Geometry of BERT. *arXiv preprint arXiv: 1906.02715*
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv: 1810.04805*
Hewitt, J., & Manning, C. D. (2019). A Structural Probe for Finding Syntax in Word Representations. Retrieved from https://nlp.stanford.edu/pubs/hewitt2019structural.pdf
Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? Retrieved from https://www.aclweb.org/anthology/P19-1356/
Kumon Institute of Education. (2010). [Table showing Kumon reading level breakdown and example exercises for levels AI-CII.] *Table of Learning Materials: Reading*. Retrieved from http://www.kumon.com/kumon_reading_levels.pdf
Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *arXiv preprint arXiv: 1611.01368*
McClellanda, J. L., Hill, F., Rudolphc, M., Baldridged, J., & Schützee Hinrich. (2019). Extending Machine Language Models toward Human-Level. *arXiv preprint arXiv: 1912.05877*