Practical Report

Devashish Kamble

1 Introduction

In recent years, various pre-trained language models that are based on the Transformers [10] architecture, in both its auto-regressive (models that use their own output as input to next time-steps and that process tokens from left-to-right, like GPT-3 [1]) and denoising (models trained by corrupting/masking the input and that process tokens bidirectionally, like BERT [3]) variants, have found to be useful in different natural language processing(NLP) tasks. These language models, having trained on enourmous volumes of text, acquire broad knowledge about the world, thus achieving high performance on a wide range of NLP benchmarks.

Even though the large neural network models appear to produce impressive performance on a range of NLP tasks, they are notoriously referred to as "black boxes". Without an understanding of how such models work behind the scenes with respect to the representations and algorithms underlying their linguistic reasoning, it is difficult to translate their success into new theories or insights about language itself, keeping in mind the relevance to linguistics and cognitive science. In order to understand the working of these models under the hood, researchers have come up with different strategies and one of them falls on the line of inquiring what linguistic knowledge is contained in these models.

Several methods have been proposed and implemented to obtain meaningful explanations and to understand how these models are able to capture the syntax and semantic-sensitive phenomena. The methods can be classified into different buckets based on the type of approach; Behavioural tests ([4] [11]), Probing Tasks ([5], [8]), Analysis of attention mechanisms [2] and Feature Attribution methods [9]. This report aims to replicate the Behavioural tests performed by Yoav Goldberg in his tech report *Assessing BERT's Syntactic Abilities* and compare the performance of five language models belonging to the BERT family [6] (bert-base-uncased, distilbert-base-uncased, roberta-base, bert-large-uncased and robert-large) on the **subject-verb agreement** grammar rule.

2 Data

The dataset utilised in this experiment is the one defined in "Targeted Syntactic Evaluation of Language Models" [7] which consists of minimally different pairs of English sentences, each being either grammatical or ungrammatical. The sentence pairs represent different variations of structure-sensitive phenomena: subject-verb agreement, reflexive anaphora and negative polarity items. The dataset is available publicly and can be downloaded from `https://github.com/yoavg/bert-syntax`.

3 Model

The Huggingface Transformers library is used to carry out the experiment as it provides APIs and tools to easily download and train state-of-the-art pretrained models. For this specific scenario, I utilise the Pipeline object (`https://huggingface.co/docs/transformers/main_classes/pipelines`), which offers a simple API dedicated to several tasks (e.g. Named Entity Recognition, Masked Language Modeling, Feature Extraction, and more). The pipeline() object can be instantiated as follows:

$$nlp = pipeline(< task_name >, model =< model_name >)$$

The models used in the experiment are as follows:

1. BERT-Base (bert-base-uncased): It is a bidirectional transformer pre-trained using a combination of masked language modelling objective and next sentence prediction on a large corpus comprising the Toronto Book Corpus and Wikipedia. This way, the model learns an inner representation of the English language that can then be used to extract features useful for downstream tasks.

2. BERT-Large (bert-large-uncased): It is an extension to the Base model by boosting the model parameters.

3. DistilBERT (distilbert-base-uncased): A transformers model, smaller and faster than BERT, which was pretrained on the same corpus in a self-supervised fashion, using the BERT base model as a teacher. As this model is uncased, it does not make a difference between english and English.

4. RoBERTa-Base (roberta-base): It builds on BERT and modifies key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates. This model being case-sensitive makes a difference between english and English.

5. RoBERTa-Large (roberta-large): It is an extension to the Base model by boosting the model parameters.

Model	**Layers**	**Hidden**	**Attention Heads**	**Parameters**
bert-base-uncased	12	768	12	110
bert-large-uncased	24	1024	16	340
distilbert-base-uncased	6	768	12	66
roberta-base	12	768	12	125
roberta-large	24	1024	16	355

Table 1: Details of the models

4 Experimental Setup

While there are many types of linguistic knowledge that one may want to investigate, one of the linguistic phenomena that provides a strong basis for analysis is the subject-verb agreement grammar rule in English, which requires that the grammatical number of a verb agree with that of the subject. For example, the sentence *"The cheetahs run."* is grammatical because *"cheetahs"* and *"run"* are both plural, but *"The cheetahs runs."* is ungrammatical because *"runs"* is a singular verb.

Linzen and Marvin came up with the Targeted Syntactic Evaluation (TSE) framework for assessing the linguistic knowledge of a language model, in which minimally different pairs of sentences, one grammatical and one ungrammatical, are shown to a model, and the model determines which one is grammatical. TSE can be used to test knowledge of the English subject–verb agreement rule by asking the model to judge between two versions of the same sentence: one where a particular verb is written in its singular form, and the other in which the verb is written in its plural form.

Goldberg used TSE to measure English subject–verb agreement ability in a BERT model. In this setup, BERT performs a fill-in-the-blank task (e.g *"the game that the guard hates [MASK] bad"*) by assigning probabilities to both the singular and plural forms of a given verb (e.g., *"is"* and *"are"*). If the model has correctly learned to apply the subject-verb agreement rule, then it should consistently assign higher probabilities to the verb forms that make the sentences grammatically correct. The task is then extended onto the different BERT-based models mentioned before.

As part of analysing the models under the subject-verb agreement rule the following three task are considered:

1. Simple agreement: It consists of minimal pair illustrating the fact that third person present English verbs agree with the number of their subject. For example,

(a) *The author laughs.*

(b) **The author laugh.*

The LM would be fed the first two words of the sentence, and would be considered successful on the task if it predicts $P(laughs) > P(laugh)$.

2. Short VP coordination: In verb phrase (VP) coordination, both of the verbs need to agree with the subject. For example,

   (a) *The senator smiles and laughs.*

   (b) **The senator smiles and laugh.*

   There are both singular and plural subjects. The number of the verb immediately adjacent to the subject is always grammatical.

3. Long VP coordination: It consists of coordination condition with a longer dependency.
   *The manager writes in a journal every day and likes/*like to watch television shows.*

After loading the dataset and the model, we can simple call the pipeline on one item (i.e. sentence) as follows:

$$predictions = nlp(< sentence >, targets =< target_tokens >)$$

The targets parameters allows us to provide the model with a set of target tokens in order to compute their probability for the MLM task. As a final step, the model output is evaluated and results are presented in the next section. The experiment can be extended/replicated by using the code present in the repo at https://github.com/devashishk99/Syntactic-abilities-BERT-family

5 Results

In order to compute the performance of the model, I have considered accuracy as a metric. Specifically, I verify how many times the models were able to assign a higher probability to the target word with the correct subject-verb agreement. There are a few stimuli in which the focus verb or its plural/singular inflection does not appear as a single word in the vocabulary and hence cannot be predicted by the model. In case of RoBERTa, there is a loss of 20 instances for the simple agreement task, 120 instances for short VP agreement task and 200 instances for long VP agreement task. The table below displays the accuracy of the models for the different subject-verb agreement task:

Model	*Simple Agreement*	*Short VP coordination*	*Long VP coordination*
bert-base-uncased	1.0	0.89	0.98
bert-large-uncased	1.0	0.86	0.98
distilbert-base-uncased	1.0	0.93	0.94
roberta-base	0.97	0.91	1.0
roberta-large	0.98	0.94	0.97
human	0.96	0.82	0.82

Table 2: Model Performance

6 Discussion

The different BERT models are trained to predict words that receive just a list of words as their input and it does not explicitly represent syntax. The high performance ($> 85\%$ atleast) of such models to solve linguistic tasks like subject-verb agreement provide evidence to the hypothesis that the models are able to learn the hierarchical structure of language. In order to perform an English subject-verb agreement we require an understanding of hierarchical structure. In the sentences *The chef is here* and *The chefs are here* the form of the verb depends on whether the subject is singular or

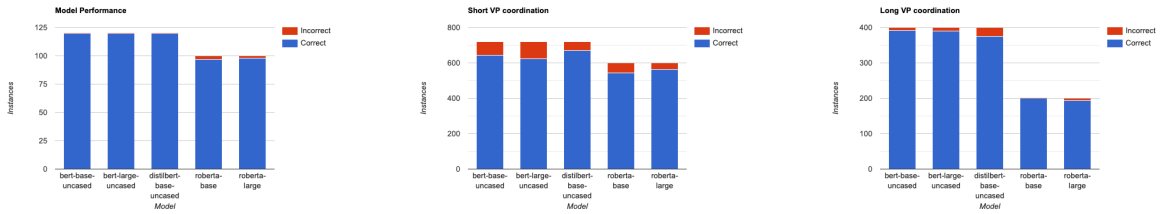

Figure 1: Model Accuracy in the tasks

plural: *is* agrees with *chef* but *are* does not. The neural language models can easily learn statistics about sequential co-occurrence i.e a singular noun is often followed by a singular verb and that is no big feat for them. However, subject-verb agreement is a special task that is not based on the linear ordering of the words, but on the words' syntactic relationship. For example, in *The chef who made the pizzas is here*, the intervening phrase does not affect the correct agreement of *is* despite the phrase containing an 'attractor' noun *pizzas*, which has the opposite number to the verb's subject.

One might have a hard time digesting the fact that these models are just too good as compared to humans on seeing the difference in performance, but the numbers for human performance come from the study by Marvin and Linzen and a direct comparison of the two cannot be made as the experimental setup is not the same.

Even though RoBERTa is trained on the MLM objective and has a larger vocabulary than BERT, it fails to outperform in the simple agreement task but then shines bright in the short and long VP coordination tasks, so much as to have almost full accuracy. DistilBERT is the model worth-noting as it has 40% less parameters than bert-base-uncased but also runs 60% faster and has an overall best performance in all the tasks.

Given the speed at which newer, faster and better models are coming up everyday it would be interesting to see how they perform on such various tasks in order to understand the linguistic capabilities they hold in their inner workings. The models put to test in this experiment were trained on English datasets which are available in plenty, but the real feat would be to replicate such results for low resource languages. As a future study, multilingual models (mBERT) should be evaluated to figure out if and how they capture rich morphology, word order and other linguistic phenomena of different languages.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[2] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert's attention, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[4] Yoav Goldberg. Assessing bert's syntactic abilities, 2019.

[5] John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short*

Papers), pages 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] Huggingface. Huggingface models. https://huggingface.co/models, 2022.

[7] Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models, 2018.

[8] Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure, 2020.

[9] Sahana Ramnath, Preksha Nema, Deep Sahni, and Mitesh M. Khapra. Towards interpreting BERT for reading comprehension based QA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3236–3242, Online, November 2020. Association for Computational Linguistics.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[11] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.