

Assessing BERT’s Syntactic Abilities

Yoav Goldberg^{1,2}

¹ Computer Science Department, Bar Ilan University

² Allen Institute for Artificial Intelligence

yogo@cs.biu.ac.il , yoav@allenai.org

Abstract

I assess the extent to which the recently introduced BERT model captures English syntactic phenomena, using (1) naturally-occurring subject-verb agreement stimuli; (2) “colorless green ideas” subject-verb agreement stimuli, in which content words in natural sentences are randomly replaced with words sharing the same part-of-speech and inflection; and (3) manually crafted stimuli for subject-verb agreement and reflexive anaphora phenomena. The BERT model performs remarkably well on all cases.

1 Introduction

The recently introduced BERT model (Devlin et al., 2018) exhibits strong performance on several language understanding benchmarks. To what extent does it capture syntax-sensitive structures?

Recent work examines the extent to which RNN-based models capture syntax-sensitive phenomena that are traditionally taken as evidence for the existence in hierarchical structure. In particular, in (Linzen et al., 2016) we assess the ability of LSTMs to learn subject-verb agreement patterns in English, and evaluate on naturally occurring wikipedia sentences. (Gulordava et al., 2018) also consider subject-verb agreement, but in a “colorless green ideas” setting in which content words in naturally occurring sentences are replaced with random words with the same part-of-speech and inflection, thus ensuring a focus on syntax rather than on selectional-preferences based cues. Marvin and Linzen (2018) consider a wider range of syntactic phenomena (subject-verb agreement, reflexive anaphora, negative polarity items) using manually constructed stimuli, allowing for greater coverage and control than in the naturally occurring setting.

The BERT model is based on the “Transformer” architecture (Vaswani et al., 2017), which—in

contrast to RNNs—relies purely on attention mechanisms, and does not have an explicit notion of word order beyond marking each word with its absolute-position embedding. This reliance on attention may lead one¹ to expect decreased performance on syntax-sensitive tasks compared to RNN (LSTM) models that do model word order directly, and explicitly track states across the sentence. Indeed, Tran et al. (2018) finds that transformer-based models perform worse than LSTM models on the Linzen et al. (2016) agreement prediction dataset. In contrast, (Tang et al., 2018) find that self-attention performs on par with LSTM for syntax sensitive dependencies in the context of machine-translation, and performance on syntactic tasks is correlated with the number of attention heads in multi-head attention.

I adapt the evaluation protocol and stimuli of Linzen et al. (2016), Gulordava et al. (2018) and Marvin and Linzen (2018) to the bidirectional setting required by BERT, and evaluate the pre-trained BERT models (both the LARGE and the BASE models). Surprisingly (at least to me), the out-of-the-box models (without any task-specific fine-tuning) perform very well on all the syntactic tasks.

2 Methodology

I use the stimuli provided by (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018), but change the experimental protocol to adapt it to the bidirectional nature of the BERT model. This requires discarding some of the stimuli, as described below. Thus, the numbers are not strictly comparable to those reported in previous work.

¹Indeed, they led me.

2.1 Previous setups

All three previous work use uni-directional language-model-like models.

Linzen et al. (2016) start with existing sentences from wikipedia that contain a present-tense verb. They feed each sentence word by word into an LSTM, stop right before the focus verb, and ask the model to predict a binary plural/singular decision (supervised setup) or compare the probability assigned by a pre-trained language model (LM) to the plural vs singular forms of the verb (LM setup).² The evaluation is then performed on sentences with “agreement attractors” in which at there is at least one noun between the verb and its subject, and all of the nouns between the verb and subject are of the opposite number from the subject.

Gulordava et al. (2018) also start with existing sentences. However, in order to control for the possibility of the model learning to rely on “semantic” selectional-preferences cues rather than syntactic ones, they replace each content word with random words from the same part-of-speech and inflection. This results in “coloreless green ideas” nonce sentences. The evaluation is then performed similarly to the LM setup of Linzen et al. (2016): the sentence is fed into a pre-trained LSTM LM up to the focus verb, and the model is considered correct if the probability assigned to the correct inflection of the original verb form given the prefix is larger than that assigned to the incorrect inflection.

Marvin and Linzen (2018) focus on manually constructed and controlled stimuli, that also emphasizes linguistic structure over selectional preferences. They construct minimal pairs of grammatical and ungrammatical sentences, feed each one *in its entirety* into a pre-trained LSTM-LM, and compare the perplexity assigned by the model to the grammatical and ungrammatical sentences. The model is “correct” if it assigns the grammatical sentence a higher probability than to the ungrammatical one. Since the minimal pairs for most phenomena differ only in a single word (the focus verb), this scoring is very similar to the one used in the two previous works. However, it does consider the continuation of the sentence after the fo-

²The results in Linzen et al. (2016) suggest that the supervised setup performs much better than the LM setup. However, several more recent work, including (Gulordava et al., 2018), show that the LM setup does get results which are very close to that of the supervised one.

cus verb, and also allows for assessing phenomena that require change into two or more words (like negative polarity items).

2.2 Adaptation to the BERT model

In contrast to these works, the BERT model is bi-directional: it is trained to predict the identity of masked words based on both the prefix and suffix surrounding these words. I adapt the uni-directional setup by feeding into BERT the complete sentence, while *masking out the single focus verb*. I then ask BERT for its word predictions for the masked position, and compare the score³ assigned to the original correct verb to the score assigned to the incorrect one.

For example, for the sentence:

a 2002 systemic review of herbal products found that several herbs , including peppermint and caraway , have anti-dyspeptic effects for non-ulcer dyspepsia with “ encouraging safety profiles ” .
(from (Linzen et al., 2016))

I feed into BERT:

```
[CLS] a 2002 systemic review of herbal products found that several herbs , including peppermint and caraway , [MASK] anti-dyspeptic effects for non-ulcer dyspepsia with `` encouraging safety profiles `` . and look for the score assigned to the words have and has at the masked position.
```

Similarly, for the pair

the game that the guard hates is bad .
the game that the guard hates are bad .
(from (Marvin and Linzen, 2018)), I feed into BERT:

```
[CLS] the game that the guard hates [MASK] bad .
```

and compare the scores predicted for *is* and *are*.

This differs from Linzen et al. (2016) and Gulordava et al. (2018) by considering the entire sentence (excluding the verb) and not just its prefix leading to the verb, and differs from Marvin and Linzen (2018) by conditioning the focus verb on bidirectional context.

I use the PyTorch implementation of BERT,

³I use the pre-softmax logit scores to save a bit of computation, but this is equivalent to using probabilities.

with the pre-trained models supplied by Google.⁴ I experiment with the `bert-large-uncased` and `bert-base-uncased` models.

Discarded Material The bi-directional setup precludes using using the NPI stimuli of [Marvin and Linzen \(2018\)](#), in which the minimal pair differs in two words position, which I discard from the evaluation. I also discard the agreement cases involving the verbs *is* or *are* in [Linzen et al. \(2016\)](#) and in [Gulordava et al. \(2018\)](#), because some of them are copular construction, in which strong agreement hints can be found also on the object following the verb.⁵ This is not an issue in the manually constructed ([Marvin and Linzen, 2018](#)) stimuli due to the patterns they chose.

Finally, I discard stimuli in which the focus verb or its plural/singular inflection does not appear as a single word in the BERT word-piece-based vocabulary (and hence cannot be predicted by the model). This include discarding [Marvin and Linzen \(2018\)](#) stimuli involving the words *swims* or *admires*, resulting in 23,368 discarded pairs (out of 152,300). I similarly discard 680 sentences from ([Linzen et al., 2016](#)) where the focus verb or its inflection were one of 108 out-of-vocabulary tokens,⁶ and 28 sentence-pairs (8 tokens⁷) from ([Gulordava et al., 2018](#)).

Limitations The BERT results are not directly comparable to the numbers reported in previous work. Beyond the differences due to bidirectionality and the discarded stimuli, the BERT models are also trained on a different and larger corpus (covering both wikipedia and books).

⁴<https://github.com/huggingface/pytorch-pretrained-BERT>

⁵Results are generally a bit higher when not discarding the *is/are* cases.

⁶blames, dislike, inhabit, exclude, revolves, governs, delete, composes, overlap, edits, embrace, compose, undertakes, disagrees, redirect, persist, recognise, rotates, accompanies, attach, undertake, earn, communicates, imagine, contradicts, specialize, accuses, obtain, caters, welcomes, interprets, await, communicate, templates, qualify, reverts, achieve, achieves, govern, restricts, violate, behave, emit, contend, adopt, overlaps, reproduces, rotate, defends, submit, revolve, lend, pertain, disagree, concentrate, detects, endorses, detect, predate, persists, consume, locates, earns, predict, interact, merge, consumes, behaves, locate, predates, enhances, predicts, integrates, inhabits, satisfy, contradict, swear, activate, restrict, satisfies, redirects, excludes, violates, interacts, admires, speculate, blame, drag, qualifies, activates, criticize, assures, welcome, depart, characterizes, defend, obtains, lends, strives, accuse, recognises, characterize, contends, perceive, complain, awaits

⁷toss, spills, tosses, affirms, spill, melt, approves, affirm

| Attractors | BERT Base | BERT Large | # sents |
|------------|-----------|------------|---------|
| 1 | 0.97 | 0.97 | 24031 |
| 2 | 0.97 | 0.97 | 4414 |
| 3 | 0.96 | 0.96 | 946 |
| 4 | 0.97 | 0.96 | 254 |

Table 1: Results on the [Linzen et al. \(2016\)](#) stimuli. While not directly comparable, the numbers reported by [Linzen et al. \(2016\)](#) and [Gulordava et al. \(2018\)](#) for the 2, 3 and 4 attractor cases are substantially lower.

| Attractors | BERT Base | BERT Large | # pairs |
|------------|-----------|------------|---------|
| All | 0.83 | 0.80 | 383 |
| 0 | 0.84 | 0.80 | 311 |
| 1 | 0.81 | 0.75 | 63 |
| 2 | 0.89 | 0.89 | 9 |

Table 2: Results on the EN NONCE ([Gulordava et al., 2018](#)) stimuli. While not strictly comparable, the numbers reported by [Gulordava et al. \(2018\)](#) for the LSTM in this condition (on All) is 74.1 ± 1.6 .

Reproducibility Code is available at <https://github.com/yoavg/bert-syntax>.

3 Results

Tables 1, 2 and 3 show the results. All cases exhibit high scores—in the vast majority of the cases substantially higher than reported in previous work.⁸ As discussed above, the results are *not directly comparable to previous work*: the BERT models are trained on different (and larger) data, are allowed to access the suffix of the sentence in addition to its prefix, and are evaluated on somewhat different data due to discarding OOV items. Still, taken together, the high performance numbers indicate that the purely attention-based BERT models are likely capable of capturing the same kind of syntactic regularities that LSTM-based models are capable of capturing, at least as well as the LSTM models and probably better.

Another noticeable and interesting trend is that **larger is not necessarily better**: the BERT-Base model outperforms the BERT-Large model on many of the syntactic conditions.

⁸The only exception are the “In a sentential complement” and “Short VP coordination” conditions in Table 3. However, the much better scores of the BERT model on the other conditions in that table suggest that the high LSTM numbers on these conditions are due to overfitting to a particular construction, rather than good syntactic generalization.

| | BERT Base | BERT Large | LSTM (M&L) | Humans (M&L) | # Pairs (# M&L Pairs) |
|---|--------------|---------------|---------------|-----------------|--------------------------|
| SUBJECT-VERB AGREEMENT: | | | | | |
| Simple | 1.00 | 1.00 | 0.94 | 0.96 | 120 (140) |
| In a sentential complement | 0.83 | 0.86 | 0.99 | 0.93 | 1440 (1680) |
| Short VP coordination | 0.89 | 0.86 | 0.90 | 0.82 | 720 (840) |
| Long VP coordination | 0.98 | 0.97 | 0.61 | 0.82 | 400 (400) |
| Across a prepositional phrase | 0.85 | 0.85 | 0.57 | 0.85 | 19440 (22400) |
| Across a subject relative clause | 0.84 | 0.85 | 0.56 | 0.88 | 9600 (11200) |
| Across an object relative clause | 0.89 | 0.85 | 0.50 | 0.85 | 19680 (22400) |
| Across an object relative (no <i>that</i>) | 0.86 | 0.81 | 0.52 | 0.82 | 19680 (22400) |
| In an object relative clause | 0.95 | 0.99 | 0.84 | 0.78 | 15960 (22400) |
| In an object relative (no <i>that</i>) | 0.79 | 0.82 | 0.71 | 0.79 | 15960 (22400) |
| REFLEXIVE ANAPHORA: | | | | | |
| Simple | 0.94 | 0.92 | 0.83 | 0.96 | 280 (280) |
| In a sentential complement | 0.89 | 0.86 | 0.86 | 0.91 | 3360 (3360) |
| Across a relative clause | 0.80 | 0.76 | 0.55 | 0.87 | 22400 (22400) |

Table 3: Results on the [Marvin and Linzen \(2018\)](#) stimuli. M&L results numbers are taken from [Marvin and Linzen \(2018\)](#). The BERT and M&L numbers are *not* directly comparable, as the experimental setup differs in many ways.

4 Discussion

The BERT models perform remarkably well on all the syntactic test cases. I expected the attention-based mechanism to fail on these (compared to the LSTM-based models), and am surprised by these results. The [Gulordava et al. \(2018\)](#) and [Marvin and Linzen \(2018\)](#) conditions rule out the possibility of overly relying on selectional preference cues or memorizing the wikipedia training data, and suggest real syntactic generalization is taking place. Exploring the *extent* to which deep purely-attention-based architectures such as BERT are capable of capturing hierarchy-sensitive and syntactic dependencies—as well as the *mechanisms* by which this is achieved—is a fascinating area for future research.

References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.
- Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. [Colorless green recurrent networks dream hierarchically](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics NAACL-HLT*, pages 1195–1205.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the ability of LSTMs to learn syntax-sensitive dependencies](#). *Transactions of the Association for Computational Linguistics*, 4:521–535.
- Rebecca Marvin and Tal Linzen. 2018. [Targeted syntactic evaluation of language models](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.
- Gongbo Tang, Mathias Mller, Annette Rios, and Rico Sennrich. 2018. [Why self-attention? a targeted evaluation of neural machine translation architectures](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4263–4272.
- Ke Tran, Arianna Bisazza, and Christof Monz. 2018. [The importance of being recurrent for modeling hierarchical structure](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008.