

Synth-tax: Modification of BERT Syntax Representations via Structural Probes

Stanford CS224N Custom Project
Mentor: John Hewitt; No external collaborators or project sharing

Kai-Siang Ang
kaiang@stanford.edu

Julia Gong
jxgong@stanford.edu

Benjamin Newman
blnewman@stanford.edu

Abstract

Large, pretrained, ‘black-box’ language models like BERT have recently risen to prominence, and we seek to understand how they represent linguistic information by causally relating models’ predictions to their internal representations. In particular, we provide a careful framework of assumptions and causality, and we manipulate BERT’s purported sentence-syntactic representations provided by the structural probe [1] to examine whether the resulting predictions of subject-verb number agreement changes. While our evidence is inconclusive, we find that BERT likely does not encode syntax in exactly the manner structural probe’s success suggests, either capturing syntax poorly or also conflating it with other linguistic information like number. In addition, we find evidence that BERT performs worse on subject-verb number agreement tasks when words are broken into separate wordpieces.

1 Introduction

Large pretrained models that produce contextual word embeddings have increasingly dominated natural language processing; these word embeddings (or some fine-tuned version of them) are effective starting points for downstream tasks like text classification, text generation, summarization, question-answering, and many more. The widespread success of these models suggests that the learned embeddings encode many important aspects of language.

In particular, these tasks require knowledge of syntax, like in question-answering where one must be able to distinguish the subject of a question. While pre-neural NLP often included explicit grammatical rules to solve these tasks, modern NLP techniques are not as interpretable, so researchers have asked, “What exactly do modern models know about syntax? How do they encode it?”

Two main approaches try to answer these questions. First, in the *behavioral* approach, researchers hand-craft model inputs and analyze the changes in model predictions. For example, to see if a model has a notion of number agreement, one might vary the number of a subject and record how much probability mass is assigned to a singular verb versus a plural one. Work in this vein has analyzed many linguistic phenomena including subject-verb agreement, filler dependencies, and anaphor agreement [2, 3, 4, 5, 6]. Second, the *probing* approach focuses on what information can be extracted from a model’s internal representations, usually by training shallow classifiers called “probes” to predict linguistic aspects like gender directly from the model representations [7, 8].

Both approaches investigate *what* is represented. [1] recently began to address the *how* by asking whether, under some linear transformation, model representations encode syntax, with the squared Euclidean distance of words’ transformed representations matching their distances in the dependency tree. They are indeed able to find such a linear transformation which constitutes their “structural probe,” suggesting that certain Euclidean subspaces of certain layers of BERT represent syntax.

However, it is unknown if the features probes find are actually used by the models. In this project, we hand-craft changes to BERT’s representations in the structural probe’s subspace (simply, “the

subspace”) and analyze BERT’s predictions, thus combining the previous approaches and seeking to draw *causal* conclusions, not just *correlative* ones. We try to address two questions:

Q1 Does the subspace meaningfully encode syntax, i.e. in a way informing model predictions?

Q2 Does that subspace avoid encoding other components of language?

We focus on the task of subject-verb number agreement in sentences with distractor nouns, such as

The author that eats many sandwiches [MASK] full. (Ex 1)

The masked verb must be singular to agree with the subject “author,” but a model that only looks at the noun closest to the verb may predict that the verb is plural due to the distractor noun “sandwiches.”

Previous work has shown that BERT and other pretrained models perform decently at this task [9, 10, 11]. If the structural probe’s subspace encodes syntax, we expect certain modifications to representations in this subspace—like moving the distractor close to the verb—to hurt BERT’s performance, but we find little such degradation. Due to the confounding possibility that the subspace conflates syntax with number, we are unable to definitively answer either one of **Q1** or **Q2** separately. Nevertheless, we know at least one answer must be “No,” i.e. the structural probe’s subspace fails to exactly encode syntax. Separately, we find that BERT-base-cased performs better than its uncased version, but it struggles when tokens are split into separate wordpieces.

2 Related Work

Previous work has similarly attempted to unite the behavioral and probing analyses. One of them uses probes trained to predict the number of a verb to modify hidden representations, changing the probability assigned to the true verb versus the one with opposite number [12]. Additionally, [13] investigate this problem in the context of neural machine translation and are able to find and control the number and gender of translated words. Though they provide insight into the encoding of number, neither addresses the existence of a “syntactic subspace” that the structural probe suggests.

Other work focuses on pretrained models’ performance at a subject-verb number agreement task in the presence of syntactically distracting information. [2] provide a dataset of these phenomena, including relative clauses, prepositional phrases, and negative polarity items. [10, 11] run BERT and OpenAI’s GPT-2 on this data showing that these models are able to perform very well on these syntactic agreement tasks. We use this dataset and provide ways to augment its vocabulary, and we validate our approach against these previous results.

3 Approach

The causal nature of our approach is inspired by [13]. Some prior works like theirs design meaningful modifications to internal model activations, but none do so to make claims about the organization of the information they manipulate. By targeting a specific subspace, we test a specific hypothesis of how BERT represents syntax. For a messy history of our thinking, see Appendix A.

3.1 General Scheme

Consider a sentence like Ex 1: it has a subject, a masked main verb (referred to as simply “the verb”), a distractor noun, and no other nouns. The subject and distractor have opposite number. Running this sentence up to some intermediate layer in BERT yields hidden representations $\vec{s}_h, \vec{v}_h, \vec{d}_h$ for the subject, verb, and distractor, respectively. The structural probe specifies a matrix B that transforms hidden representations $\vec{s}_h, \vec{v}_h, \vec{d}_h$ into their *syntactic representations*

$$\vec{s} := B\vec{s}_h, \quad \vec{v} := B\vec{v}_h, \quad \vec{d} := B\vec{d}_h.$$

Now, we intervene with a potentially non-deterministic function f , replacing the distractor’s syntactic representation $\vec{d}$ with $\vec{d}' := f(\vec{d})$. This causes the distractor’s hidden representation to change into

$$\vec{d}'_h := \vec{d}_h - B^+ \vec{d} + B^+ \vec{d}',$$

where B^+ is the pseudo-inverse of B . We let BERT continue from this layer using modified hidden representations $\vec{s}_h, \vec{v}_h, \vec{d}_h$, resulting in a probability distribution P over BERT’s vocabulary that represents its predictions for the masked verb. Finally, we measure some metric $M(P) \in [0, 1]$ of this prediction’s incorrectness, i.e. how much it agrees with the distractor instead of the subject.

In the following subsections, we consider this scheme with various f applied to one common sentence.

3.2 Assumptions

Given that previous work indicates BERT understands number (Section 2), we assume:

- A1** A hidden representations $\vec{x}_h$ encodes number via some function $N_h(\vec{x}_h) \in [0, 1]$, where 0 means agreement with the subject and 1 means agreement with the distractor.
- A2** BERT understands the subject and distractor’s numbers, i.e. $N_h(\vec{s}_h) = 0, N_h(\vec{d}_h) = 1$.
- A3’** BERT predicts the verb’s number according to its perception of the numbers of the nouns in the sentence, i.e. $M(P)$ is an increasing function in both $N_h(\vec{s}_h)$ and $N_h(\vec{d}_h)$.

Our consideration of the syntax subspace and **Q1** induces the following:

- A3** Some parameter $\alpha \in [0, 1]$ describes the degree to which BERT looks at the distractor instead of the subject, so we amend **A3’** to instead be that $M(P)$ is an increasing function in both $(1 - \alpha)N_h(\vec{s}_h)$ and $\alpha N_h(\vec{d}_h)$.

We also consider two confounders:

- A4** In answering **Q2**, we consider the possibility that the subspace encodes number, i.e. a syntactic representation $\vec{x}$ encodes number via some function $N(\vec{x}) \in [0, 1]$ (same convention as **A1**). Similarly, a representation’s non-syntactic component $\vec{x}_\perp := \vec{x}_h - B^+ \vec{x}$ encodes number via some function $N_\perp(\vec{x}_\perp) \in [0, 1]$. Notice that $\vec{d}'_\perp = \vec{d}_\perp$. Moreover, we assume the number $N_h(\vec{x}_h)$ of the overall hidden representation is a non-negatively weighted average of the numbers $N_\perp(\vec{x}_\perp), N(\vec{x})$ of its components, i.e.

$$N_h(\vec{x}_h) = \gamma(\vec{x}_h)N(\vec{x}) + (1 - \gamma(\vec{x}_h))N_\perp(\vec{x}_\perp), \quad \gamma(\vec{x}_h) \in [0, 1],$$

with $\gamma(\vec{x}_h) = 0$ if the subspace does not encode number, i.e. the answer to **Q2** is “Yes.”

- A5** The perturbation of the distractor’s syntactic representation moves the representation off of BERT’s typical manifold, degrading performance. We specifically assume $M(P)$ is an increasing function of $\delta := \|\vec{d}' - \vec{d}\|$, but alternatives like $\|\vec{d}'_h - \vec{d}_h\|$ are also reasonable.¹

Finally, we assume simplicity:

- A6’** There are no other factors impacting $M(P)$. Using the previous assumptions, we can deduce a redescription of $M(P)$ as a function $\widetilde{M}$, increasing in all its arguments:

$$M(P) = \widetilde{M}\left(\alpha, \delta, -N_h(\vec{s}_h), N_\perp(\vec{d}'_\perp), N(\vec{d}')\right).$$

In Section 3.4, we will see we need to expand these assumptions.

3.3 Causality

We know that $N_h(\vec{s}_h) = 0$ due to **A2**, and that $N_\perp(\vec{d}'_\perp) = N_\perp(\vec{d}_\perp)$ is constant under our interventions, hence by **A6’**, $M(P)$ is an increasing function of α, δ , and $N(\vec{d}')$, with no other factors.

In our interventions, we hypothesize that certain interventions f will induce a $\vec{d}'$ that makes BERT look at the distractor more or less, i.e. manipulate α . The above analysis allows us to draw select conclusions about α . Consider comparing these statistics between a *reference intervention* f_0 to a *targeted intervention* f_1 , with subscripts 0 and 1 appended to their associated vectors and values:

¹Our choice allows us to assume that $\vec{d}'_\perp$ remains unchanged under baselines that try to account for δ .

If $M(P_0) > M(P_1)$, $\delta_0 \leq \delta_1$, and $N(\vec{d}_0) \leq N(\vec{d}_1)$, then it must certainly be true that $\alpha_0 > \alpha_1$.

Similarly, if $M(P_0) < M(P_1)$, $\delta_0 \geq \delta_1$, and $N(\vec{d}_0) \geq N(\vec{d}_1)$, then $\alpha_0 < \alpha_1$.

In plain English, BERT is more wrong when we perturb more, when we make the distractor’s number more unlike the subject, and when BERT looks at the distractor more. If BERT becomes more right even though we perturb at least the same amount and make the distractor’s number at least as unlike the subject, then BERT has to be looking at the distractor more. A similar conclusion can be made for when BERT becomes less right.

Lastly, note that if the answer to **Q2** is “Yes,” then the number $N(\vec{d}')$ in the syntax component does not exist, so as long as we match δ s, we can conclude that $M(P)$ and α move in the same way.

3.4 Reference Interventions

When do the conditions in Section 3.3 occur? Though each intervention can precisely control δ , only **A2** explains an intervention’s effect on $N(\vec{d}')$, yielding a select few reference interventions f_0 :

- **Match subject:** Let f_0 move the distractor exactly to the subject, so $N(\vec{d}_0) = N(\vec{s}) = 0$, giving $N(\vec{d}_0) \leq N(\vec{d}_1)$. If $\delta_0 \leq \delta_1$ and $M(P_0) > M(P_1)$, then $\alpha_0 > \alpha_1$.
- **Naive baseline:** If f_0 is the identity, $N(\vec{d}_0)$ takes on the maximal value $N(\vec{d})$ and $\delta_0 = 0$, giving $N(\vec{d}_0) \geq N(\vec{d}_1)$, $\delta_0 \leq \delta_1$. Ignoring the unverifiable case of $N(\vec{d}_0) = N(\vec{d}_1)$, we are unable to make any conclusions from this reference intervention.
- **Baseline:** To make use of the approach in the naive baseline, we need larger δ_0 , but $N(\vec{d}_0)$ should remain large. The natural thing to do is to average $\vec{d}_0$ over uniformly random δ_0 -perturbations around $\vec{d}$, but there’s no guarantee that this averages out to a large $\mathbb{E}[N(\vec{d}_0)]$, since $N(\vec{d})$ was already maximal. Hence, in order to conclude that $\mathbb{E}[\alpha_0] < \alpha_1$ if we find some f_1 with $\mathbb{E}[M(P_0)] < M(P_1)$ and $\delta_0 \geq \delta_1$, we must assume the following:

A6 Technically, the expectation values require us to assume linearity of $\widetilde{M}$, amending **A6’**. Letting $\Delta(\delta) \in [0, 1]$ be an increasing function of δ with $\Delta(0) = 0$ and $\lim_{\delta \rightarrow \infty} \Delta(\delta) = 1$:

$$M(P) = \Delta(\delta) + (1 - \Delta(\delta)) \left[(1 - \alpha)N_h(\vec{s}_h) + \alpha \left(\gamma(\vec{d}_h)N_\perp(\vec{d}_\perp) + (1 - \gamma(\vec{d}_h))N(\vec{d}) \right) \right].$$

A7 The distractor’s number under f_1 is no more than its average under f_0 : $\mathbb{E}[N(\vec{d}_0)] \geq N(\vec{d}_1)$.

These are both strong assumptions, but **A6** is probably reasonable as some first-order approximation, and **A7** holds in some plausible scenarios: imagine dividing the syntax subspace between “closer to the subject/verb” and “closer to the distractor,” with the former region having lower $N(\vec{d})$ than the second; our targeted interventions should move the distractor to the former region to satisfy **A7**.

3.5 Syntax Subspace Hypotheses

We hypothesize four possible ways that BERT might determine how much it looks at the distractor:

1. **Absolute/relative distance** from verb: BERT looks at the distractor more when it’s close to the verb in absolute distance, or it’s closer to the verb than the subject.
2. **Absolute location** in subspace: BERT looks at the distractor more when it’s close to some special location $\vec{z}$ in the subspace. For simplicity, we assume $\vec{z} = \vec{s}$, similar to **A2**.
3. **Location relative to subject-verb** line: BERT looks at the distractor more when it’s around some special line from the verb to $\vec{z}$, and for simplicity, we assume $\vec{z} = \vec{s}$, again like **A2**. This could mean that BERT looks at the distractor more when it’s close to the verb or close to the line, or when the distractor-verb- $\vec{z}$ angle is small.
4. **Location relative to distractor-verb** line: BERT looks at the distractor less when it’s around some special line from the verb to $\vec{z}$, and for simplicity, we assume $\vec{z} = \vec{d}$, again like **A2**. This could mean that BERT looks at the distractor more when it’s close to the verb or away from the line, or when the distractor-verb- $\vec{z}$ angle is large.

Mathematically, when each of these holds, we expect see the following in our results:

1. **Absolute/relative distance:** α is decreasing in $\|\vec{d}' - \vec{v}\|$ or $\|\vec{d}' - \vec{v}\|/\|\vec{s} - \vec{v}\|$.
2. **Absolute location:** α is decreasing in $\|\vec{d}' - \vec{s}\|$.
3. **Location relative to subject-verb:** Let ℓ_s be the line through $\vec{s}$ and $\vec{v}$, and let $\vec{p}'_s$ be the projection of $\vec{d}'$ onto ℓ_s . α is decreasing in $\|\vec{p}'_s - \vec{d}'\|$, $\|\vec{p}'_s - \vec{d}'\|/\|\vec{p}'_s - \vec{v}\|$, and $\|\vec{p}'_s - \vec{v}\|$.
4. **Location relative to distractor-verb:** Let ℓ_d be the line through $\vec{d}$, $\vec{v}$, and let $\vec{p}'_d$ be $\vec{d}'$'s projection onto ℓ_d . α is increasing in $\|\vec{p}'_d - \vec{d}'\|$, $\|\vec{p}'_d - \vec{d}'\|/\|\vec{p}'_d - \vec{v}\|$, decreasing in $\|\vec{p}'_d - \vec{v}\|$.

3.6 Targeted Interventions

We design four corresponding targeted interventions. Let r be some real number, let $\hat{\cdot}$ denote vector normalization, let $\text{proj}_x(y)$ be the projection of y onto x , let $\mathcal{S}$ denote a sample from the unit sphere, and let k be the minimum of 1 and a sample from an exponential distribution with rate parameter λ .²

1. **Verb-spherical:** $f_1(\vec{d}) = \vec{v} + r\mathcal{S}$, $r \geq 0$; Samples $\vec{d}'$ at radius r around the verb.
2. **Subject-spherical:** $f_1(\vec{d}) = \vec{s} + r\mathcal{S}$, $r \geq 0$; Samples $\vec{d}'$ at radius r around the subject.
3. **Subject-verb-conical:** $f_1(\vec{d}) = \vec{v} + r\widehat{(\vec{s} - \vec{v})} + kr\widehat{\vec{q}}$ where $\vec{q} = \vec{z} - \text{proj}_{\vec{s}-\vec{v}}(\vec{z})$, $\vec{z} = \mathcal{S}$; Samples $\vec{p}'_s$ at distance r from the verb, then moves orthogonally away from ℓ_s to $\vec{d}'$ such that $k = \|\vec{p}'_s - \vec{d}'\|/\|\vec{p}'_s - \vec{v}\|$. One can think of $\vec{v}$ as the cone's vertex and $\vec{s}$ as being on the cone's axis of symmetry. We choose k to concentrate samples near this axis.
4. **Distractor-verb-conical:** $f_1(\vec{d}) = \vec{v} + r\widehat{(\vec{d} - \vec{v})} + kr\widehat{\vec{q}}$ where $\vec{q} = \vec{z} - \text{proj}_{\vec{d}-\vec{v}}(\vec{z})$, $\vec{z} = \mathcal{S}$; Similar to the above subject-verb-conical.

For reference, here are the mathematical forms of the reference interventions from Section 3.4.

Match subject: $f_0(\vec{d}) = \vec{s}$, **Naive baseline:** $f_0(\vec{d}) = \vec{d}$, **Baseline:** $f_0(\vec{d}) = \vec{d} + r\mathcal{S}$, $r \geq 0$.

In each intervention, we record $M(P)$ and the quantities relevant to the hypotheses in Section 3.5: pairwise distances between $\vec{s}$, $\vec{v}$, $\vec{d}$, and $\vec{d}'$, between $\vec{d}'$ and $\vec{p}'_s$ or $\vec{p}'_d$, and between $\vec{v}$ and $\vec{p}'_s$ or $\vec{p}'_d$.³

4 Experiments

4.1 Data

We use data from [2], containing minimal pairs of procedurally generated sentences. We focus on 5,600 sentences with intervening relative clauses and 8,030 with intervening prepositional phrases with animate and inanimate subjects/verbs; subject and distractor always have opposite number. We replace main verbs with [MASK] tokens, and to ease aligning these sentences to the structural probe, we filter out sentences with tokens not in BERT's wordpiece vocabulary. Finally, for each sentence, we append a period and capitalize the first letter to match cased-BERT's training data.

4.2 Evaluation Method

We propose three specific binary metrics to measure BERT's error $M(P)$, so that 0 means right and that 1 means wrong. These provide alternative definitions right and wrong.

1. **Single Disagreement (SD):** One "correct" verb is supplied with each sentence, so compare the probability mass BERT puts on the correct verb to its form with the opposite number. There are only three distinct verb pairs in the dataset (is/are, smiles/smile, laughs/laugh).

²Actually, we use $\sqrt{r}$, $\sqrt{k}$ in place of r , k because the *squared* Euclidean distances in the syntax subspace are designed to match the dependency tree distances.

³Since $\vec{p}'_s$, $\vec{p}'_d$ are on lines through $\vec{v}$, they have a notion of signed distance from $\vec{v}$, calculable via dot products. Also, we modify all distances to represent the tree distances, which involves squaring the Euclidean distance.

2. **Probability Mass Vote (PMV)**: Some verbs in the procedurally generated data are uncommon and might be out-of-domain for BERT. We take a list of the 500 most common verbs in English [14], filter them down to 491 non-modal verbs in BERT’s vocabulary, and sum up the probability mass assigned to all their singular forms compared to the mass on their plural forms. (is/are and smiles/smile are on this list; laughs/laugh is not.)
3. **Equal Verb Vote (EVV)**: Same as the probably mass vote, except instead of comparing the sum of probability masses on singular versus plural verbs, we compare each pair of verbs and give that pair a vote assigned to whichever form has more mass; majority wins. This may ameliorate the case where one verb pair is assigned a lot more mass than the others.

4.3 Experimental Details

Our experiments follow the scheme in Section 3.1. Our structural probe is a rank-128 distance probe trained for layer 7 of BERT-base-cased on the Penn Treebank dataset, as done in [1]. We obtain our hidden states from BERT-base-cased layer 7, and we generate predictions using the masked language-modeling head adapted from the Huggingface transformers library implementation [15].

We present seven intervention types in Section 3.6. For those dependent on r , we let r take on 16 values linearly spaced between r_{low} and r_{high} , inclusive. For the baseline, spherical f_1 , and conical f_1 , $(r_{\text{low}}, r_{\text{high}})$ are $(0.5, 9)$, $(0.5, 5)$, $(-2, 4)$, respectively. Since these f involve randomness, for each f and value of r , we take 16 random samples. The two remaining f involve neither r nor sampling.

Our code at github.com/bnewm0609/causal-probes-p has design based on github.com/john-hewitt/structural-probes with some copied code in our `src/{probe, run}.py`.

5 Results and Analysis

5.1 Reproducing Prior Work

In attempting to reproduce the results of [10] by running the naive baseline, i.e. no intervention, we found significant differences due to two factors. First, we use BERT-base-cased, while they use the uncased version. Second, we filter out sentences with any words not in BERT’s wordpiece vocabulary; they only do so when the main verb is not in BERT’s wordpiece vocabulary.⁴

Table 3 of [10] reports one minus the SD metric averaged over all sentences. In Table 1, we show our reproduction of their results (Line 4). Moreover, changing their set-up (Line 4) with either one of our filtering or our version of BERT (Lines 3 and 2, respectively), reduces error a bit. Changing both factors to our set-up (Line 1) reduces error very significantly.

Line #	Data Filtering Method	Bert Model	SD	PMV	# of sentences
1	Ours	base-cased	0.04	0.02	16210
2	[10]	base-cased	0.12	0.08	19440
3	Ours	base-uncased	0.12	0.07	16210
4	[10]	base-uncased	0.15	0.10	19440

Table 1: Comparison of our set-up to [10]. We reproduce their reported error (SD) in Line 4. Line 1 (bold) is this paper’s general configuration.

We conclude that BERT-base-cased seems significantly better at this number agreement task than the uncased version, but it struggles dramatically with words outside its wordpiece vocabulary.

5.2 Qualitative Analysis of Hypotheses

We now examine correlations between various collected measures and the error $M(P)$ across all datasets (subject relative, prepositional animate, prepositional inanimate). In our plots, each point represents an average across all sentences and all samples of a given radius r , along with the associated

⁴Unlike the rest of the paper, the data here also includes sentences where the subject and distractor share the same number, to match [10].

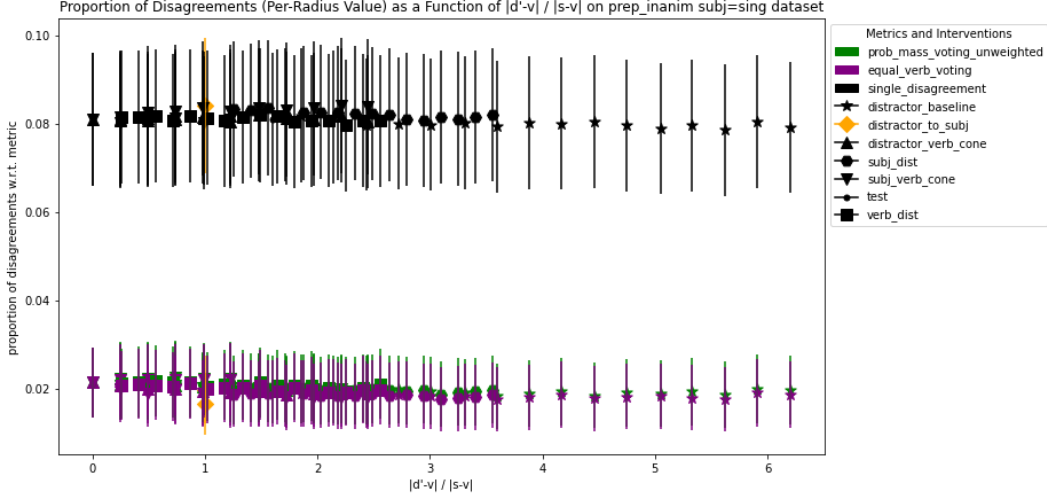


Figure 1: Error $M(P)$ vs. $||\vec{d}' - \vec{v}'|| / ||\vec{s} - \vec{v}'||$ on the prepositional inanimate, singular subject dataset.

95% confidence interval. Orange indicates the match subject intervention, and other colors indicate the choice of $M(P)$ (see Section 4.2). The shapes indicate other interventions.

We consider each hypothesis from Section 3.5, checking if we see what we expect if they were true:

0. **Absolute distance:** We find little or mildly positive correlation between $M(P)$ and $||\vec{d}' - \vec{v}'||$, so the evidence is inconclusive as to whether BERT uses absolute distance.
1. **Relative distance:** We find slight positive correlation between $M(P)$ and $||\vec{d}' - \vec{v}'|| / ||\vec{s} - \vec{v}'||$ for most datasets, an inconclusive result. In the preposition inanimate dataset, we find slight negative correlation (Figure 1). This suggests BERT uses relative distance for that dataset.
2. **Absolute location:** In the subject relative and preposition animate datasets, we find $M(P)$ positively correlates with $||\vec{d}' - \vec{s}'||$, an inconclusive result. In the preposition inanimate dataset, we find a bowl-shaped graph for interventions other than the baseline (Figure 2). This suggests BERT uses absolute location for that dataset, at least in some region around $\vec{s}'$.
3. **Location relative to subject-verb:** We find little or mildly positive correlation between $M(P)$ and $||\vec{p}'_s - \vec{d}'||$, $||\vec{p}'_s - \vec{d}'|| / ||\vec{p}'_s - \vec{v}'||$, and $||\vec{p}'_s - \vec{v}'||$, which is an inconclusive result.
4. **Location relative to distractor-verb:** We find little correlation between $M(P)$ and $||\vec{p}'_d - \vec{d}'||$, $||\vec{p}'_d - \vec{d}'|| / ||\vec{p}'_d - \vec{v}'||$, and $||\vec{p}'_d - \vec{v}'||$. This is inconclusive.

We find that large perturbations degrade performance (Figure 3), validating **A4**. However, we find no other strong trends. This tells us that the syntax subspace is imperfect: perturbations in it do not seem to have any expected interpretable effect, either because syntax is not well-encoded in this subspace or because this space also encodes number. Thus, at least one answer to **Q1** or **Q2** is “No.”

5.3 Problematic Words

We examine when BERT fails by finding words that are predictive of failure by calculating the pointwise mutual information (PMI) between a random variable (RV) representing BERT’s error and an RV of the probability that a sentence contains a particular word. Across sentences with subject-modifying prepositional phrases, we find that the words “teacher(s),” “smile(s),” “taxi driver(s),” “skater(s),” and “song(s)” are most predictive of BERT making an error. For sentences with subject-modifying relative clauses, we find similar words. The most problematic word here is “smile(s),” the main verb of the sentence, suggesting that BERT’s performance often degrades due to out-of-domain error, not number agreement error. See Appendix B for more results.

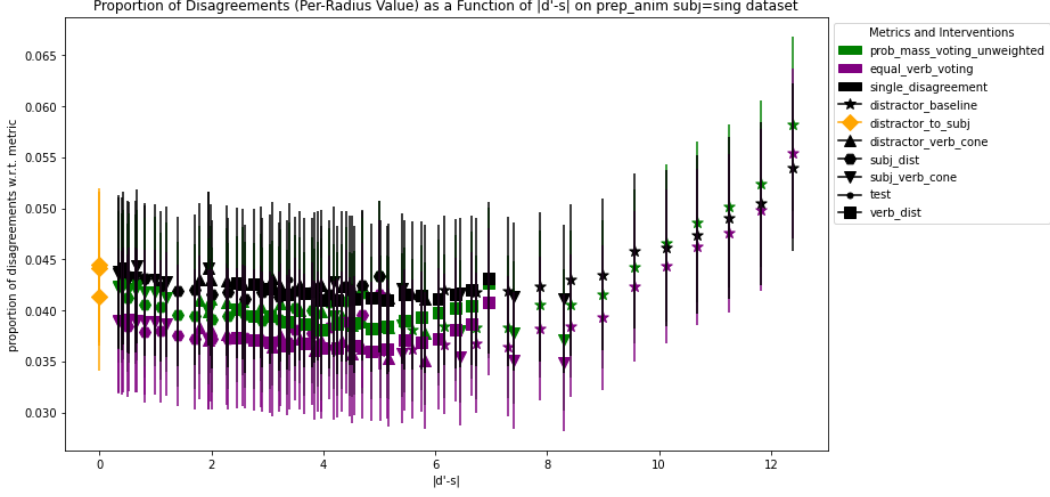


Figure 2: Error $M(P)$ vs. $\|\vec{d}' - \vec{s}\|$ on the prepositional animate, singular subject dataset.

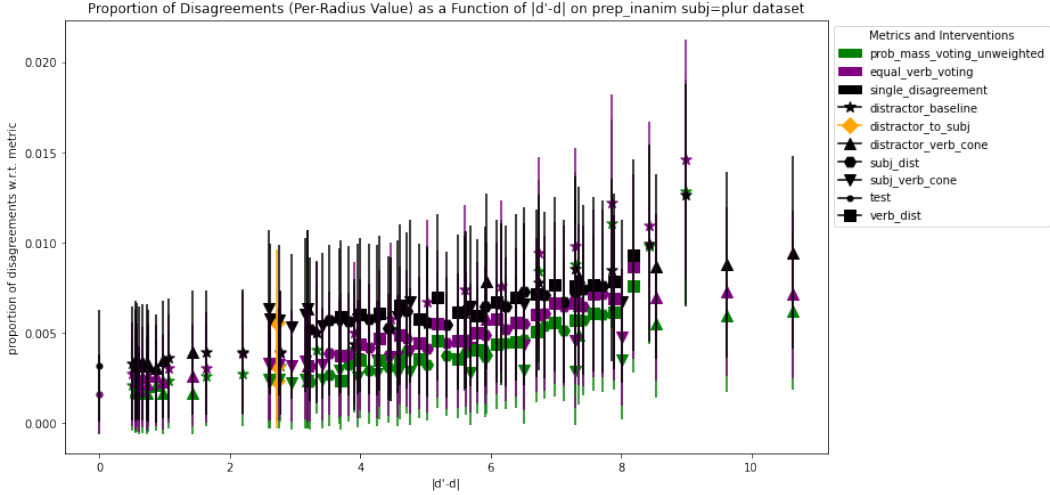


Figure 3: Error $M(P)$ vs. $\|\vec{d}' - \vec{d}\|$ on the preposition inanimate, plural subject dataset.

6 Conclusion

While our results do not point to BERT representing syntax in a particular way, they do illustrate that the subspace does not represent syntax exactly, either failing to fully capture syntax or accidentally capturing other linguistic features as well. In the future, we plan on taking into account how other work has probed BERT to investigate its ability to perform well at syntactic phenomena and see if we can make causal inferences based on their approaches. In addition, our formal causal framework was under-used and under-validated; we would like to formally test the assumptions underlying the analysis, and reduce how many there are if possible, while using the theoretic conclusions we derive to make conclusions about BERT. Finally, subject-verb number agreement is not the only task that requires syntactic knowledge; tasks like prepositional phrase attachment and gender agreement could let us peer further into how BERT represents syntax.

References

- [1] 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, 2019.
- [2] Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1192–1202, Brussels, Belgium, 2018. Association for Computational Linguistics.
- [3] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics (TACL)*, 4:521–535, 2016.
- [4] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [5] Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*, 2018.
- [6] Jennifer Hu, Sherry Y Chen, and Roger P Levy. A closer look at the performance of neural language models on reflexive anaphor licensing. *Proceedings of the Society for Computation in Linguistics*, 3(1):382–392, 2020.
- [7] Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *ACL*, 2018.
- [8] Dieuwke Hupkes and Willem Zuidema. Visualisation and ‘diagnostic classifiers’ reveal how recurrent and recursive neural networks process hierarchical structure (extended abstract). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5617–5621. International Joint Conferences on Artificial Intelligence Organization, 7 2018.
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [10] Yoav Goldberg. Assessing bert’s syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019.
- [11] Thomas Wolf. Some additional experiments extending the tech report “assessing bert’s syntactic abilities” by yoav goldberg. Technical report, Technical report, 2019.
- [12] Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium, November 2018. Association for Computational Linguistics.
- [13] Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Identifying and controlling important neurons in neural machine translation. *arXiv preprint arXiv:1811.01157*, 2018.
- [14] Mark Davies. The corpus of contemporary american english (coca): 600 million words, 1990-present. 2008.
- [15] Thomas Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv, abs/1910.03771*, 2019.

A Appendix: History of Approach

Our mentor and project TA, John Hewitt, recommended that our report include not just the final product, but also evidence of the messy history for how we arrived there. To that end, the interested reader can flip through this 12-page Overleaf: <https://www.overleaf.com/read/dgcbkskzzttp>.

It contains the chronological musings that led to our final experimental design.

Section 1.1 sets up a simple mathematical formalism that lays our assumptions bare. Section 1.2 proposes some targeted interventions. Section 1.3 thinks about what can be concluded from the interventions under simplifying assumptions. Section 1.4 decides we need to take into account more factors, and it approximately thinks about causality. Section 1.5 contains a summary of these thoughts, and Section 1.6 points to the future work at that point.

Section 2.1 tries to generalize the targeted interventions. Section 2.2 thinks about hypotheses for how the syntax subspace works. Section 2.3 realizes issues with causality. Section 2.4 properly addresses causality. Section 2.5 returns to interventions that explore particular hypotheses about the syntax subspace. Section 2.6 lays out a coding spec sheet for the targeted interventions. Section 2.7 summarizes what we learned.

Section 3 contains notes for debugging our intervention code.

B Appendix: Words PMI

B.1 Calculation

Given a binary random variable, D which represents whether or not BERT disagrees or doesn't disagree with a sentence, and a set of binary random variables $\mathbf{W}$, where W_i represents the probability of a word i appearing in a sentence, we calculate the pointwise mutual information between D and W_i for each W_i as:

$$\text{PMI}(D, W_i) = \log \frac{P(D | W_i)}{P(D)} \quad (1)$$

We estimate $P(D)$ using an MLE estimate, as the ratio between the number of sentences where BERT disagrees and the total number of sentences. To estimate $P(D | W_i)$ for each of the 39 words in the Marvin and Linzen vocabulary, we use Laplace smoothing. We calculate the disagreements using both the singular disagreement metric and the probability mass voting metric.

B.1.1 Results

Here we report the PMI between words and disagreement, broken down by the Marvin and Linzen categories: Prepositions with animate objects, inanimate objects, and subject-modifying relative clauses. Each with singular subjects and plural subjects:

Singular Subject				Plural Subject			
Word	PMI (SD)	Word	PMI (PMV)	Word	PMI (SD)	Word	PMI (PMV)
smiles	1.13	smiles	0.94	assistant	1.50	skater	2.01
laughs	0.94	laughs	0.94	smile	1.28	pilots	1.42
author	0.40	officer	0.69	teachers	1.24	farmers	1.42
teacher	0.38	chefs	0.69	pilots	1.09	smile	1.25
pilot	0.38	ministers	0.65	officers	1.09	laugh	1.25
surgeon	0.36	pilot	0.62	guard	1.09	like	1.10
officer	0.36	skaters	0.58	skater	0.91	surgeons	0.91
assistants	0.34	parents	0.58	like	0.87	parent	0.91
parents	0.30	teacher	0.46	laugh	0.74	officers	0.91
skaters	0.25	surgeon	0.37	parent	0.68	assistant	0.91
ministers	0.21	author	0.32	hate	0.47	hate	0.51
farmer	0.21	assistants	0.32	surgeons	0.40	the	0.08
chefs	0.18	loves	0.10	farmers	0.40	that	0.08
loves	0.11	the	0.01	the	0.05	teachers	-0.19
the	0.00	that	0.01	that	0.05	taxi	-0.19
that	0.00	hates	-0.00	young	-0.11	senators	-0.19
hates	-0.03	likes	-0.05	love	-0.51	minister	-0.19
likes	-0.07	farmer	-0.11	old	-0.52	managers	-0.19
customer	-0.14	young	-0.14	taxi	-0.70	guard	-0.19
executives	-0.25	customer	-0.19	senators	-0.70	executive	-0.19
architects	-0.25	architects	-0.48	minister	-0.70	driver	-0.19
dancers	-0.29	manager	-0.73	managers	-0.70	dancer	-0.19
manager	-0.33	executives	-0.73	executive	-0.70	customers	-0.19
guards	-0.33	dancers	-1.07	driver	-0.70	consultants	-0.19
taxi	-0.37	is	-1.43	dancer	-0.70	chef	-0.19
drivers	-0.37	guards	-1.58	customers	-0.70	authors	-0.19
young	-1.02	old	-2.09	consultants	-0.70	architect	-0.19
is	-2.40	taxi	-2.68	chef	-0.70	young	-0.70
senator	-3.51	senator	-2.68	authors	-0.70	tall	-0.70
consultant	-3.51	drivers	-2.68	architect	-0.70	short	-0.70
tall	-4.01	consultant	-2.68	are	-1.21	old	-0.70
short	-4.01	tall	-3.18	tall	-1.21	love	-1.10
old	-4.01	short	-3.18	short	-1.21	admire	-1.10

Table 2: Subject-relative clauses. SD = single disagreement. PMV = probability mass voting. Positive values are more predictive of BERT choosing the wrong verb number, while negative values are more predictive of BERT choosing the correct verb number.

Singular Subject				Plural Subject			
Word	PMI (SD)	Word	PMI (PMV)	Word	PMI (SD)	Word	PMI (PMV)
author	1.33	author	1.18	teachers	1.86	skater	2.06
smiles	1.05	smiles	0.94	skater	1.30	teachers	1.60
laughs	1.01	laughs	0.94	smile	1.20	near	1.35
officer	0.77	officer	0.85	taxi	0.79	authors	1.27
parents	0.70	parents	0.70	driver	0.79	guard	1.04
assistants	0.61	assistants	0.61	in	0.63	in	0.94
ministers	0.47	ministers	0.51	front	0.63	front	0.94
next	0.43	pilot	0.46	pilots	0.61	smile	0.76
chefs	0.31	chefs	0.41	guard	0.61	laugh	0.76
behind	0.23	next	0.31	of	0.51	pilots	0.76
to	0.21	behind	0.27	side	0.44	officers	0.76
skaters	0.18	skaters	0.16	dancer	0.39	tall	0.54
pilot	0.18	teacher	0.08	behind	0.28	of	0.40
near	0.06	near	0.05	chef	0.25	dancer	0.35
the	0.01	to	0.03	authors	0.25	young	0.25
side	-0.04	the	0.01	near	0.10	the	0.07
teacher	-0.13	in	-0.05	the	0.02	side	-0.16
surgeon	-0.13	front	-0.05	tall	-0.00	old	-0.16
of	-0.17	from	-0.16	to	-0.12	are	-0.29
farmer	-0.22	across	-0.16	laugh	-0.41	taxi	-0.34
in	-0.26	surgeon	-0.18	are	-0.54	surgeons	-0.34
front	-0.26	of	-0.19	young	-0.59	senators	-0.34
from	-0.39	side	-0.28	short	-0.59	parent	-0.34
across	-0.39	taxi	-0.28	old	-0.59	minister	-0.34
taxi	-0.44	farmer	-0.28	parent	-0.59	managers	-0.34
guards	-0.44	drivers	-0.28	officers	-0.59	farmers	-0.34
drivers	-0.44	young	-0.43	from	-0.81	executive	-0.34
dancers	-0.44	dancers	-0.54	across	-0.81	driver	-0.34
young	-0.73	executives	-0.69	managers	-1.00	customers	-0.34
manager	-0.73	manager	-0.87	farmers	-1.00	consultants	-0.34
executives	-0.73	architects	-0.87	customers	-1.00	chef	-0.34
customer	-0.91	old	-1.04	next	-1.10	assistant	-0.34
architects	-1.14	guards	-1.10	surgeons	-1.69	architect	-0.34
is	-2.02	customer	-1.10	senators	-1.69	to	-0.85
old	-2.34	is	-1.38	minister	-1.69	short	-0.85
senator	-2.52	tall	-2.30	executive	-1.69	next	-0.85
consultant	-2.52	senator	-2.48	consultants	-1.69	from	-0.85
tall	-3.03	consultant	-2.48	assistant	-1.69	behind	-0.85
short	-3.03	short	-2.99	architect	-1.69	across	-0.85

Table 3: Intervening relative phrases with animate objects. SD = single disagreement. PMV = probability mass voting. Positive values are more predictive of BERT choosing the wrong verb number, while negative values are more predictive of BERT choosing the correct verb number.

Singular Subject				Plural Subject			
Word	PMI (SD)	Word	PMI (PMV)	Word	PMI (SD)	Word	PMI (PMV)
interests	1.94	interests	1.97	movies	2.41	minister	2.69
people	1.26	executive	1.55	to	2.16	games	2.59
parent	0.61	people	1.29	joy	2.16	interest	2.34
executive	0.55	parent	1.27	bring	2.16	people	1.65
movie	0.50	skater	1.01	taxi	1.59	taxi	1.59
skater	0.50	novel	0.91	guard	1.59	skater	1.59
show	0.27	movie	0.91	driver	1.59	parent	1.59
picture	0.27	book	0.91	dancer	1.59	guard	1.59
book	0.27	show	0.35	chef	1.59	executive	1.59
from	0.13	by	0.18	people	1.47	driver	1.59
chef	0.06	song	0.06	skater	0.90	dancer	1.59
assistant	0.06	game	0.06	parent	0.90	chef	1.59
game	0.04	the	0.04	minister	0.90	assistant	1.59
the	0.01	from	-0.04	executive	0.90	architect	1.59
minister	-0.04	taxi	-0.93	assistant	0.90	songs	1.49
architect	-0.04	minister	-0.93	architect	0.90	shows	1.49
painting	-0.04	guard	-0.93	songs	0.80	poems	1.49
novel	-0.04	driver	-0.93	shows	0.80	pictures	1.49
by	-0.11	dancer	-0.93	poems	0.80	paintings	1.49
dancer	-0.39	chef	-0.93	pictures	0.80	novels	1.49
song	-0.50	assistant	-0.93	paintings	0.80	movies	1.49
poem	-0.65	architect	-0.93	novels	0.80	books	1.49
guard	-0.73	poem	-1.04	games	0.80	unpopular	1.24
taxi	-0.95	picture	-1.04	books	0.80	to	1.24
driver	-0.95	painting	-1.04	from	0.69	popular	1.24
unpopular	-2.69	unpopular	-1.28	unpopular	0.55	new	1.24
to	-2.69	to	-1.28	popular	0.55	joy	1.24
popular	-2.69	popular	-1.28	new	0.55	good	1.24
new	-2.69	new	-1.28	interest	0.55	bring	1.24
joy	-2.69	joy	-1.28	good	0.55	bad	1.24
good	-2.69	good	-1.28	bad	0.55	from	0.69
brings	-2.69	brings	-1.28	the	0.22	by	0.69
bad	-2.69	bad	-1.28	by	-0.00	the	0.40
is	-4.29	is	-2.88	are	-1.05	are	-0.36

Table 4: Intervening relative phrases with inanimate objects. SD = single disagreement. PMV = probability mass voting. Positive values are more predictive of BERT choosing the wrong verb number, while negative values are more predictive of BERT choosing the correct verb number.