What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models

Allyson Ettinger
Department of Linguistics
University of Chicago
aettinger@uchicago.edu

Abstract

Pre-training by language modeling has become a popular and successful approach to NLP tasks, but we have yet to understand exactly what linguistic capacities these pre-training processes confer upon models. In this paper we introduce a suite of diagnostics drawn from human language experiments, which allow us to ask targeted questions about information used by language models for generating predictions in context. As a case study, we apply these diagnostics to the popular BERT model, finding that it can generally distinguish good from bad completions involving shared category or role reversal, albeit with less sensitivity than humans, and it robustly retrieves noun hypernyms, but it struggles with challenging inference and role-based event prediction—and in particular, it shows clear insensitivity to the contextual impacts of negation.

1 Introduction

Pre-training of NLP models with a language modeling objective has recently gained popularity as a precursor to task-specific fine-tuning. Pre-trained models like BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018a) have advanced the state of the art in a wide variety of tasks, suggesting that these models acquire valuable, generalizable linguistic competence during the pre-training process. However, though we have established the benefits of language model pre-training, we have yet to understand what exactly about language these models learn during that process.

This paper aims to improve our understanding of what language models (LMs) know about language, by introducing a set of diagnostics targeting a range of linguistic capacities, drawn from human psycholinguistic experiments. Because of their origin in psycholinguistics, these diagnostics have two distinct advantages: they are carefully controlled to ask targeted questions about linguistic capabilities, and they are designed to ask these questions by examining word predictions in context, which allows us to study LMs without any need for task-specific fine-tuning.

Beyond these advantages, our diagnostics distinguish themselves from existing tests for LMs in two primary ways. First, these tests have been chosen specifically for their capacity to reveal insensitivities in predictive models, as evidenced by patterns that they elicit in human brain responses. Second, each of these tests targets a set of linguistic capacities that extend beyond the primarily syntactic focus seen in existing LM diagnostics—we have tests targeting commonsense/pragmatic inference, semantic roles and event knowledge, category membership, and negation. Each of our diagnostics is set up to support tests of both word prediction accuracy and sensitivity to distinctions between good and bad context completions. Although we focus on the BERT model here as an illustrative case study, these diagnostics are applicable for testing of any language model.

This paper makes two main contributions. First, we introduce a new set of targeted diagnostics for assessing linguistic capacities in language models.[1] Second, we apply these tests to shed light on strengths and weaknesses of the popular BERT model. We find that BERT struggles with challenging commonsense/pragmatic inferences and role-based event prediction, that it is generally robust to within-category distinctions and role reversals, but with lower sensitivity than humans, and that it is very strong at associating nouns with hypernyms. Most strikingly, however, we find that BERT fails completely to show generalizable understanding of negation, raising questions about the aptitude of LMs to learn this type of meaning.

[1] All test sets and experiment code are made available here: https://github.com/aetting/lm-diagnostics

2 Related Work

This paper contributes to a growing effort to better understand the specific linguistic capacities achieved by neural NLP models. Some approaches use fine-grained classification tasks to probe information in sentence embeddings (Adi et al., 2016; Conneau et al., 2018; Ettinger et al., 2018), or token-level and other sub-sentence level information in contextual embeddings (Tenney et al., 2018; Peters et al., 2018b). Much work has attempted to evaluate systems' overall level of "understanding", often with tasks such as semantic similarity and entailment (Wang et al., 2018; Bowman et al., 2015; Agirre et al., 2012; Dagan et al., 2005; Bentivogli et al., 2016), and additional work has been done to design curated versions of these tasks to test for specific linguistic capabilities (Dasgupta et al., 2018; Poliak et al., 2018; McCoy et al., 2019). Our diagnostics complement this previous work in allowing for direct testing of language models in their natural setting—via controlled tests of word prediction in context—without requiring probing of extracted representations or task-specific fine-tuning.

Previous work has used word prediction accuracy as a test of LMs' language understanding. The LAMBADA dataset (Paperno et al., 2016), in particular, tests models' ability to predict the final word of a passage, in cases where the final sentence alone is insufficient for prediction. However, while LAMBADA presents a challenging prediction task, it is not well-suited to ask targeted questions about types of information used by LMs for prediction, as our tests are designed to do.

Some previous work does use targeted tests to examine specific capacities of LMs—often inspired by psycholinguistic methods. However, the majority of this work has focused on syntactic capabilities of LMs (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018; Wilcox et al., 2018; Futrell et al., 2019). Relevant to our case study here, using several of these tests Goldberg (2019) shows the BERT model to perform impressively on such syntactic diagnostics. In the present work, we expand to examine a more diverse range of linguistic capabilities, while continuing to use controlled, targeted diagnostics. We also deviate from previous work with targeted diagnostics in not only comparing word probabilities, but also examining word prediction accuracies directly, for a richer picture of models' specific strengths and weaknesses.

3 Leveraging human studies

The power in our diagnostics stems from their origin in psycholinguistic studies—the items have been carefully designed for studying specific aspects of language processing, and each test has been shown to produce informative patterns of results when tested on humans. In this section we provide relevant background on human language processing, and explain how we use this information to choose the particular tests used here.

3.1 Background: prediction in humans

To study language processing in humans, psycholinguists often test human responses to words in context, in order to better understand the information that our brains use to generate predictions. In particular, there are two types of predictive human responses that are relevant to us here:

Cloze probability The first measure of human expectation is a measure of the "cloze" response. In a cloze task, humans are given an incomplete sentence and tasked with filling their expected word in the blank. "Cloze probability" of a word w in context c refers to the proportion of people who choose w to complete c. We will treat this as the best available gold standard for human prediction in context—humans completing the cloze task typically are not under any time pressure, so they have the opportunity to use all available information from the context to arrive at a prediction.

N400 amplitude The second measure of human expectation is a brain response known as the N400, which is detected by measuring electrical activity at the scalp (by electroencephalography). Like cloze, the N400 can be used to gauge how expected a word w is in a context c—the amplitude of the N400 response appears to be sensitive to fit of a word in context, and has been shown to correlate with cloze in many cases (Kutas and Hillyard, 1984). The N400 has also been shown to correlate with language model probabilities (Frank et al., 2013). However, the N400 differs from cloze in being a real-time response that occurs only 400 milliseconds into the processing of a word. Accordingly, the expectations reflected in the N400 sometimes deviate from the more fully-formed expectations reflected in the untimed cloze response.

Context	Expected	Inappropriate
He complained that after she kissed him, he couldn't get the red color off his face. He finally just asked her to stop wearing that ____	*lipstick*	*mascara* | *bracelet*
He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____	*football*	*baseball* | *monopoly*

Table 1: Example items from CPRAG-34

3.2 Our diagnostic tests

The test sets that we use here are all drawn from human studies that have revealed *divergences between cloze and N400 profiles*—that is, for each of these tests, the N400 response suggests a level of insensitivity to certain information when computing expectations, causing a deviation from the fully-informed cloze predictions. We choose these as our diagnostics because they provide built-in sensitivity tests targeting the types of information that appear to have reduced effect on the N400—and because they should present particularly challenging prediction tasks, tripping up models that fail to use the full set of available information.

4 Datasets

Each of our diagnostics supports three types of testing: word prediction accuracy, sensitivity testing, and qualitative prediction analysis. Because these items are designed to draw conclusions about human processing, each set is carefully constructed to constrain the information relevant for making word predictions. This allows us to examine how well LMs use this target information.

For word prediction accuracy, we use the most expected items from human cloze probabilities as the gold completions.[2] These represent predictions that models should be able to make if they access and apply all relevant context information when generating probabilities for target words.

For sensitivity testing, we compare model probabilities for good versus bad completions—specifically on comparisons for which the N400 has exhibited reduced sensitivity in the human experiments. This allows us to test whether LMs will show similar insensitivities with respect to the relevant linguistic distinctions.

Finally, because these items are constructed in such a controlled manner, qualitative analysis of models' top predictions can be highly informative about information being applied for prediction. We leverage this in our experiments below.

In all tests, the target word to be predicted falls in the final position of the provided context, which means that these tests should function similarly for either left-to-right or bidirectional LMs. In anticipation of testing the BERT model, and to facilitate fair future comparisons with the present results, we filter out items for which the expected word is not in BERT's single-word vocabulary, to ensure that all expected words can be predicted.

It is important to acknowledge that these are small test sets, due to their origin in psycholinguistic studies. However, because these sets have been hand-designed by cognitive scientists to test predictive processing in humans, their value is in the targeted assessment that they provide with respect to information that LMs use in prediction.

We now we describe each test set in detail.

4.1 CPRAG-34: commonsense and pragmatic inference

Our first set targets commonsense and pragmatic inference, and tests sensitivity to differences within semantic category. The left column of Table 1 shows examples of these items, each of which consists of two sentences. These items come from an influential human study by Federmeier and Kutas (1999), which tested how brains would respond to different types of context completions, shown in the right columns of Table 1.

Information needed for prediction Accurate prediction on this set requires use of commonsense to infer what is being described in the first sentence, and pragmatic reasoning to determine how the second sentence relates. For instance, in Table 1, commonsense informs us that red color left by kisses suggests lipstick, and pragmatic reasoning allows us to infer that the thing to stop wear-

[2] With one exception, NEG-88, for which we use completion truth, as in the original study.

ing is related to the complaint. As in LAMBADA, the final sentence is generic, not supporting prediction on its own. Unlike LAMBADA, the consistent structure of these items allows us to target specific model capabilities,[3] and additionally, none of these items contain the target word in context,[4] forcing models to use commonsense inference rather than coreference. Human cloze probabilities show a high level of agreement on appropriate completions for these items—average cloze probability for expected completions is .74.

Sensitivity test The Federmeier and Kutas (1999) study found that while the inappropriate completions (e.g., *mascara, bracelet*) had cloze probabilities of virtually zero (average cloze .004 and .001, respectively), the N400 showed some expectation for completions that shared a semantic category with the expected completion (e.g., *mascara*, by relation to *lipstick*). Our sensitivity test targets this distinction, testing whether LMs will favor inappropriate completions based on shared semantic category with expected completions.

Data The authors of the original study make available 40 of their items—we filter out six items to accommodate BERT's single-word vocabulary, for a final set of 34 items.[5]

4.2 ROLE-88: event knowledge and semantic role sensitivity

Our second set targets event knowledge and semantic role interpretation, and tests sensitivity to impact of role reversals. Table 2 shows an example item pair from this set. These items come from a human experiment by Chow et al. (2016), which tested the brain's sensitivity to role reversals.

Information needed for prediction Accurate prediction on this set requires a model to interpret semantic roles from sentence syntax, and apply event knowledge about typical interactions between types of entities in the given roles. The set has reversals for each noun pair (shown in Table 2) so models must distinguish roles for each order.

[3]To highlight this advantage, as a supplement for this test set we provide specific annotations of each item, indicating the knowledge/reasoning required to make the prediction.

[4]More than 80% of LAMBADA items contain the target word in the preceding context.

[5]For a couple of items, we also replace an inappropriate completion with another inappropriate completion of the same semantic category to accommodate BERT's vocabulary.

Context	Compl.
the restaurant owner forgot which customer the waitress had ____	*served*
the restaurant owner forgot which waitress the customer had ____	*served*

Table 2: Example items from ROLE-88 (Compl = Context Completion)

Sensitivity test The Chow et al. (2016) study found that although each completion (e.g., *served*) is good for only one of the noun orders and not the reverse, the N400 shows a similar level of expectation for the target completions regardless of noun order. Our sensitivity test targets this distinction, testing whether LMs will show similar difficulty distinguishing appropriate continuations based on word order and semantic role. Human cloze probabilities show strong sensitivity to the role reversal, with average cloze difference of .233 between good and bad contexts for a given completion.

Data The authors provide 120 sentences (60 pairs)—which we filter to 88 final items, removing pairs for which the best completion of either context is not in BERT's single-word vocabulary.

4.3 NEG-88: negation

Our third set targets understanding of the meaning of negation, along with knowledge of category membership. Table 3 shows examples of these test items, which involve absence or presence of negation in simple sentences, with two different completions that vary in truth depending on the negation. These test items come from a human study by Fischler et al. (1983), which examined how human expectations change with the addition of negation.

Context	Match	Mismatch
A robin is a ____	*bird*	*tree*
A robin is not a ____	*bird*	*tree*

Table 3: Example NEG-88-SIMP items, with targets matching and mismatching category of subject noun

Information needed for prediction Because the negative contexts in these items are highly unconstraining (*A robin is not a ____* ?), prediction accuracy is not a useful measure for the negative contexts. We test prediction accuracy for affirmative contexts only, which allows us to test models'

use of hypernym information (*robin* = *bird*). Targeting of negation happens in the sensitivity test.

Sensitivity test The Fischler et al. (1983) study found that although the N400 shows preference for true completions in affirmative sentences (e.g., *A robin is a bird*), it fails to adjust to negation, preferring the false continuations in negative sentences (e.g., *A robin is not a bird*). Our sensitivity test targets this distinction, testing whether LMs will show similar insensitivity to impacts of negation. Note that unlike in the previous sections, here we use truth judgments rather than cloze probability as an indication of the quality of a completion.

Data Fischler et al. provide the list of 18 subject nouns and 9 category nouns that they use for their sentences, which we use to generate a comparable dataset, for a total of 72 items.[6] We refer to these 72 simple sentences as NEG-88-SIMP. All target words are in BERT's single-word vocabulary.

Supplementary items In a subsequent study, Nieuwland and Kuperberg (2008) followed up on the Fischler et al. (1983) experiment, creating affirmative and negative sentences chosen to be more "natural ... for somebody to say", and contrasting these with affirmative and negative sentences chosen to be less natural. "Natural" items include examples like *Most smokers find that quitting is (not) very (difficult/easy)*, while items designed to be less natural include examples like *Vitamins and proteins are (not) very (good/bad)*. The authors share 16 items, which we add to the 72 above for additional comparison. We refer to these supplementary 16 items, designed to test effects of naturalness, as NEG-88-NAT.

5 Experiments

As a case study, we use these three diagnostics to examine the predictive capacities of the pre-trained BERT model (Devlin et al., 2019), which has been the basis of impressive performance across a wide range of tasks. BERT is a deep bidirectional transformer network (Vaswani et al., 2017) pre-trained on tasks of masked language modeling (predicting masked words given bidirectional context) and next-sentence prediction (binary classification of whether two sentences are a sequence). We test two versions of the pre-trained model: $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ (uncased). These versions have the same basic architecture, but $\text{BERT}_{\text{LARGE}}$ has more parameters—in total, $\text{BERT}_{\text{BASE}}$ has 110M parameters, and $\text{BERT}_{\text{LARGE}}$ has 340M. We use the PyTorch BERT implementation with masked language modeling parameters for generating word predictions.[7]

For testing, we process our sentence contexts to have a [MASK] token—also used during BERT's pre-training—in the target position of interest. We then measure BERT's predictions for this [MASK] token's position. Following Goldberg (2019), we also add a [CLS] token to the start of each sentence to mimic BERT's training conditions.

BERT differs from traditional left-to-right language models, and from real-time human predictions, in being a bidirectional model able to use information from both left and right context. This difference should be neutralized by the fact that our items provide all information in the left context—however, in our experiments here, we do allow one advantage for BERT's bidirectionality: we include a period and a [SEP] token after each [MASK] token, to indicate that the target position is followed by the end of the sentence. We do this in order to give BERT the best possible chance of success, by maximizing the chance of predicting a single word rather than the start of a phrase. Items for these experiments thus appear as follows:

[*CLS*] *The restaurant owner forgot which customer the waitress had* [*MASK*] . [*SEP*]

Logits produced by the language model for the target position are softmax-transformed to obtain probabilities comparable to human cloze probability values for those target positions.[8]

6 Results for CPRAG-34

First we report BERT's results on the CPRAG-34 test targeting commonsense, pragmatic reasoning, and sensitivity within semantic category.

[6] The one modification that we make to the original concrete noun list is a substitution of the word *salmon* for *bass* within the category of fish—this was done because *bass* created lexical ambiguity in a way that was not interesting for our purposes here.

[7] PyTorch BERT: https://github.com/huggingface/pytorch-pretrained-BERT

[8] Human cloze probabilities are importantly different from true probabilities over a vocabulary, making these values not directly comparable. However, cloze provides important indication—the best indication we have—of how much a context constrains human expectations toward a continuation, so we do at times loosely compare these two types of values.

	Orig	Shuf	Trunc	Shuf + Trunc
$\text{BERT}_{\text{BASE}}$ $k = 1$	23.5	14.1 ± 3.1	14.7	8.1 ± 3.4
$\text{BERT}_{\text{LARGE}}$ $k = 1$	35.3	17.4 ± 3.5	17.6	10.0 ± 3.0
$\text{BERT}_{\text{BASE}}$ $k = 5$	52.9	36.1 ± 2.8	35.3	22.1 ± 3.2
$\text{BERT}_{\text{LARGE}}$ $k = 5$	52.9	39.2 ± 3.9	32.4	21.3 ± 3.7

Table 4: CPRAG-34 word prediction accuracies (with and without sentence perturbations)

6.1 Word prediction accuracies

We define accuracy as percentage of items for which the "expected" completion is among the model's top k predictions, with $k = 1$ and $k = 5$.

Table 4 ("Orig") shows the accuracies of $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$. For accuracy at $k = 1$, $\text{BERT}_{\text{LARGE}}$ soundly outperforms $\text{BERT}_{\text{BASE}}$ with correct predictions on just over a third of items. When we expand to $k = 5$, the models converge on the same accuracy, identifying the expected completion in approximately half of items.

Because commonsense and pragmatic reasoning are non-trivial concepts to pin down, it is worth asking to what extent BERT can achieve this performance based on simpler cues like word identities or n-gram context. To test importance of word order, we shuffle the words in each item's first sentence, garbling the message but leaving all individual words intact ("Shuf" in Table 4). To test adequacy of truncated context, we remove all words of the second sentence but the two words preceding the target word ("Trunc"). This gives generally enough syntactic context to identify the part of speech, as well as some sense of semantic category, but removes other information from that second sentence. We also test with both perturbations together ("Shuf + Trunc"). Because different shuffled word orders give rise to different results, for the "Shuf" and "Shuf + Trunc" settings we show mean and standard deviation from 100 runs.

Table 4 shows the accuracies as a result of these perturbations. One thing that is immediately clear is that the BERT model is indeed making use of information provided by the word order of the first sentence, and by the more distant content of the second sentence, as each of these individual perturbations causes a notable drop in accuracy. It is worth noting, however, that with each perturbation there is a subset of items for which BERT's accuracy remains intact. Unsurprisingly, many of these items are those containing particularly distinctive words associated with the target, such as *checkmate* (*chess*), *touchdown* (*football*), and *stone-washed* (*jeans*). This suggests that some of BERT's success on these items may be attributable to simpler lexical or n-gram information. In Section 6.3 we take a closer look at some more difficult items that seemingly avoid such loopholes.

6.2 Completion sensitivity

Next we test BERT's ability to prefer expected completions over inappropriate completions of the same semantic category. We first test this by simply measuring the percentage of items for which BERT assigns a higher probability to the good completion (e.g., *lipstick* from Table 1) than to either of the inappropriate completions (e.g., *mascara, bracelet*). Table 5 shows the results. We see that $\text{BERT}_{\text{BASE}}$ assigns the highest probability to the expected completion in 73.5% of items, while $\text{BERT}_{\text{LARGE}}$ does so for 79.4%—a solid majority, but with a clear portion of items for which an inappropriate, semantically-related target does receive a higher probability than the appropriate word.

	Prefer good	w/ .01 thresh
$\text{BERT}_{\text{BASE}}$	73.5	44.1
$\text{BERT}_{\text{LARGE}}$	79.4	58.8

Table 5: Percent of CPRAG-34 items with good completion assigned higher probability than bad

We can make our criterion slightly more stringent if we introduce a threshold on the probability difference. The average cloze difference between good and bad completions is about .74 for the data from which these items originate, reflecting a very strong human sensitivity to the difference in completion quality. To test the proportion of items in which BERT assigns more substantially different probabilities, we filter to items for which the good completion probability is higher by greater than .01—a threshold chosen to be very generous given the significant average cloze difference. With this threshold, the sensitivity drops

Context	$\text{BERT}_{\text{LARGE}}$ predictions
Pablo wanted to cut the lumber he had bought to make some shelves. He asked his neighbor if he could borrow her ____	*car, house, room, truck, apartment*
The snow had piled up on the drive so high that they couldn't get the car out. When Albert woke up, his father handed him a ____	*note, letter, gun, blanket, newspaper*
At the zoo, my sister asked if they painted the black and white stripes on the animal. I explained to her that they were natural features of a ____	*cat, person, human, bird, species*

Table 6: $\text{BERT}_{\text{LARGE}}$ predictions on selected CPRAG-34 items

noticeably—$\text{BERT}_{\text{BASE}}$ shows sensitivity in only 44.1% of items, and $\text{BERT}_{\text{LARGE}}$ shows sensitivity in only 58.8%. These results tell us that although the models are able to prefer good completions to same-category bad completions in a majority of these items, the difference is in many cases very small, suggesting that this sensitivity falls short of what we see in human cloze responses.

6.3 Qualitative examination of predictions

We see above that the BERT models are able to identify the correct word completions in approximately half of CPRAG-34 items, and that the models are able to prefer good completions to semantically-related inappropriate completions in a majority of items, though with notably weaker sensitivity than humans. To better understand the models' weaknesses, in this section we examine predictions made when the models fail.

Table 6 shows three example items along with the top five predictions of $\text{BERT}_{\text{LARGE}}$. In each case, BERT provides completions that are sensible in the context of the second sentence, but that fail to take into account the context provided by the first sentence—in particular, the predictions show no evidence of having been able to infer the relevant information about the situation or object described in the first sentence. For instance, we see in the first example that BERT has correctly zeroed in on things that one might borrow, but it fails to infer that the thing to be borrowed is something to be used for cutting lumber. Similarly, BERT's failure to detect the snow-shoveling theme of the second item makes for an amusing set of non sequitur completions. Finally, the third example shows that BERT has identified an animal theme (unsurprising, given the words *zoo* and *animal*), but it is not applying the phrase *black and white stripes* to identify the appropriate completion of *zebra*. Altogether, these examples illustrate that with respect to the target capacities of commonsense inference and pragmatic reasoning, BERT fails in these more challenging cases.

7 Results for ROLE-88

Next we turn to the ROLE-88 test of semantic role sensitivity and event knowledge.

7.1 BERT prediction accuracy

We again define accuracy by presence of a top cloze item within the model's top k predictions. Table 8 ("Orig") shows the accuracies for $\text{BERT}_{\text{LARGE}}$ and $\text{BERT}_{\text{BASE}}$. For $k = 1$, accuracies are very low, with $\text{BERT}_{\text{BASE}}$ slightly outperforming $\text{BERT}_{\text{LARGE}}$. When we expand to $k = 5$, accuracies predictably increase, and $\text{BERT}_{\text{LARGE}}$ now outperforms $\text{BERT}_{\text{BASE}}$ by a healthy margin.

To test the extent to which BERT is relying on the individual nouns in the context, we try two different perturbations of the contexts: removing the information from the object (*which customer the waitress ...*), and removing the information from the subject (*which customer the waitress...*), in each case by replacing the noun with a generic substitute. In order to make the nouns highly generic, we choose *one* and *other* as the substitutions for the object and subject, respectively.

Table 8 shows the results with each of these perturbations individually and together. We observe several notable patterns. First, removing either the object ("-Obj") or the subject ("-Subj") has relatively little effect on the accuracy of $\text{BERT}_{\text{BASE}}$ for either $k = 1$ or $k = 5$. This is quite different from what we see with $\text{BERT}_{\text{LARGE}}$, the accuracy of

Context	$\text{BERT}_{\text{BASE}}$ predictions	$\text{BERT}_{\text{LARGE}}$ predictions
the camper reported which girl the bear had ____	*taken, killed, attacked, bitten, picked*	*attacked, killed, eaten, taken, targeted*
the camper reported which bear the girl had ____	*taken, killed, fallen, bitten, jumped*	*taken, left, entered, found, chosen*
the restaurant owner forgot which customer the waitress had ____	*served, hired, brought, been, taken*	*served, been, delivered, mentioned, brought*
the restaurant owner forgot which waitress the customer had ____	*served, been, chosen, ordered, hired*	*served, chosen, called, ordered, been*

Table 7: $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ predictions on selected ROLE-88 sentences

	Orig	-Obj	-Sub	-Both
$\text{BERT}_{\text{BASE}}$ k=1	14.8	12.5	12.5	9.1
$\text{BERT}_{\text{LARGE}}$ k=1	13.6	5.7	6.8	4.5
$\text{BERT}_{\text{BASE}}$ k=5	27.3	26.1	22.7	18.2
$\text{BERT}_{\text{LARGE}}$ k=5	37.5	18.2	21.6	14.8

Table 8: ROLE-88 word prediction accuracies (with and without sentence perturbations)

	$\leq$.17	$\leq$.23	$\leq$.33	$\leq$.77
$\text{BERT}_{\text{BASE}}$ k=1	12.0	17.4	17.4	11.8
$\text{BERT}_{\text{LARGE}}$ k=1	8.0	4.3	17.4	29.4
$\text{BERT}_{\text{BASE}}$ k=5	24.0	26.1	21.7	41.1
$\text{BERT}_{\text{LARGE}}$ k=5	28.0	34.8	39.1	52.9

Table 9: Accuracy of predictions in unperturbed ROLE-88 sentences, by max cloze bins

which drops substantially when the object or subject information is removed. These patterns suggest that $\text{BERT}_{\text{BASE}}$ is less dependent upon the full detail of the subject-object structure, instead relying primarily upon one or the other of the participating nouns for its verb predictions. $\text{BERT}_{\text{LARGE}}$, on the other hand, appears to make heavier use of both nouns, such that loss of either one causes non-trivial disruption in the predictive accuracy.

It should be noted that the items in this set are overall less constraining than those in Section 6—humans converge less clearly on the same predictions, resulting in lower average cloze values for the best completions. To investigate the effect of constraint level, we divide items into four bins by top cloze value per sentence. Table 9 shows the results. With the exception of $\text{BERT}_{\text{BASE}}$ at $k = 1$, for which accuracy in all bins is fairly low, it is clear that the highest cloze bin yields much higher model accuracies than the other three bins, suggesting some alignment between how constraining contexts are for humans and how constraining they are for BERT. However, even in the highest cloze bin, when at least a third of humans converge on the same completion, even $\text{BERT}_{\text{LARGE}}$ at $k = 5$ is only correct in half of cases, suggesting substantial room for improvement.[9]

	Prefer good	w/ .01 thresh
$\text{BERT}_{\text{BASE}}$	75.0	31.8
$\text{BERT}_{\text{LARGE}}$	86.4	43.2

Table 10: Percent of ROLE-88 items with good completion assigned higher probability than role reversal

7.2 Completion sensitivity

Next we test BERT's sensitivity to role reversals by comparing model probabilities for a given completion (e.g., *served*) in the appropriate versus role-reversed contexts. We again start by testing the percentage of items for which BERT assigns a higher probability to the appropriate than to the inappropriate completion. As we see in Table 10, $\text{BERT}_{\text{BASE}}$ prefers the good continuation in 75% of items, while $\text{BERT}_{\text{LARGE}}$ does so for 86.4%—comparable to the proportions for CPRAG-34. However, when we apply our threshold of .01—still generous given the average cloze difference of .233—sensitivity drops more dramatically than on CPRAG-34, to 31.8% and 43.2%.

Overall, these results suggest that BERT is, in a majority of cases of this kind, able to use noun position to prefer good verb completions to bad—however, it is again less sensitive than humans to these distinctions, and it fails to match human

[9]This analysis is made possible by the authors' generous provision of the cloze data for these items, which was not originally made public with the items themselves.

word predictions on a strong majority of cases. The model's ability to choose good completions over role reversals (albeit with weak sensitivity) suggests that the failures on word prediction accuracy are not due to inability to distinguish word orders, but rather to a weakness in event knowledge or understanding of semantic role implications.

7.3 Qualitative examination of predictions

Table 7 shows predictions of $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ for some illustrative examples. For the *girl*/*bear* items, we see that $\text{BERT}_{\text{BASE}}$ favors continuations like *killed* and *bitten* with *bear* as subject, but also includes these continuations with *girl* as subject. $\text{BERT}_{\text{LARGE}}$, by contrast, excludes these continuations when *girl* is the subject.

In the second pair of sentences we see that the models choose *served* as the top continuation under both word orders, even though for the second word order this produces an unlikely scenario. In both cases, the model's assigned probability for *served* is much higher for the appropriate word order than the inappropriate one—a difference of .6 for $\text{BERT}_{\text{LARGE}}$ and .37 for $\text{BERT}_{\text{BASE}}$—but it is noteworthy that no more semantically appropriate continuation is identified by either model for *which waitress the customer had* ____.

As a final note, although the continuations are generally impressively grammatical, we see exceptions in the second *bear*/*girl* sentence—both models produce completions of questionable grammaticality (or at least questionable use of selection restrictions), with sentences like *which bear the girl had fallen* from $\text{BERT}_{\text{BASE}}$, and *which bear the girl had entered* from $\text{BERT}_{\text{LARGE}}$.

8 Results for NEG-88

Finally, we turn to the NEG-88 test of negation and category membership.

8.1 BERT prediction accuracy

We start by testing the ability of BERT to predict correct category continuations for the affirmative contexts in NEG-88-SIMP. Table 11 shows the accuracy results for these affirmative sentences.

We see that for $k = 5$, the correct category is predicted for 100% of affirmative items, suggesting an impressive ability of both BERT models to associate nouns with their correct immediate hypernyms. We also see that the accuracy drops substantially when assessed on $k = 1$. Examination of predictions reveals that these errors consist exclusively of cases where BERT completes the sentence with a repetition of the subject noun, e.g., *A daisy is a daisy*—which is certainly true, but which is not a likely or informative sentence.

	Accuracy
$\text{BERT}_{\text{BASE}}$ $k = 1$	38.9
$\text{BERT}_{\text{LARGE}}$ $k = 1$	44.4
$\text{BERT}_{\text{BASE}}$ $k = 5$	100
$\text{BERT}_{\text{LARGE}}$ $k = 5$	100

Table 11: Accuracy of word predictions in NEG-88-SIMP affirmative sentences

8.2 Completion sensitivity

We next assess BERT's sensitivity to the meaning of negation, by measuring the proportion of items in which the model assigns higher probabilities to true completions than to false ones.

	Affirmative	Negative
$\text{BERT}_{\text{BASE}}$	100	0.0
$\text{BERT}_{\text{LARGE}}$	100	0.0

Table 12: Percent of NEG-88-SIMP items with true completion assigned higher probability than false

Table 12 shows the results, and the pattern is stark. When the statement is affirmative (*A robin is a* ____), the models assign higher probability to the true completion in 100% of items. Even with the threshold of .01—which eliminated many comparisons on CPRAG-34 and ROLE-88—all items pass but one (for $\text{BERT}_{\text{BASE}}$), suggesting a robust preference for the true completions.

However, in the negative statements (*A robin is not a* ____), BERT prefers the true completion in 0% of items, assigning the higher probability to the false completion in every case. This shows a strong insensitivity to the meaning of negation, with BERT preferring the category match completion every time, despite its falsity.

8.3 Qualitative examination of predictions

Table 13 shows examples of the predictions made by $\text{BERT}_{\text{LARGE}}$ in positive and negative contexts. We see a clear illustration of the phenomenon suggested by the results above: for affirmative sentences, BERT produces generally true completions (at least in the top two)—but these completions remain largely unchanged after negation is added,

Context	$\text{BERT}_{\text{LARGE}}$ predictions
A robin is a ____	*bird, robin, person, hunter, pigeon*
A daisy is a ____	*daisy, rose, flower, berry, tree*
A hammer is a ____	*hammer, tool, weapon, nail, device*
A hammer is an ____	*object, instrument, axe, implement, explosive*
A robin is not a ____	*robin, bird, penguin, man, fly*
A daisy is not a ____	*daisy, rose, flower, lily, cherry*
A hammer is not a ____	*hammer, weapon, tool, gun, rock*
A hammer is not an ____	*object, instrument, axe, animal, artifact*

Table 13: NEG-88-SIMP predictions by $\text{BERT}_{\text{LARGE}}$

resulting in many blatantly untrue completions.

Another interesting phenomenon that we can observe in Table 13 is BERT's sensitivity to the nature of the determiner (*a* or *an*) preceding the masked word. This determiner varies depending on whether the upcoming target begins with a vowel or a consonant (for instance, our mismatched category paired with *hammer* is *insect*) and so the model can potentially use this cue to filter the predictions to those starting with either vowels or consonants. How effectively does BERT use this cue? The predictions indicate that BERT is for the most part extremely good at using this cue to limit to words that begin with the right type of letter. There are certain exceptions (e.g., *An ant is not a ant*), but these are in the minority.

8.4 Increasing naturalness

The supplementary NEG-88-NAT items allow us to examine further the model's handling of negation, with items designed to test the effect of "naturalness". When we present BERT with this new set of sentences, the model does show an apparent change in sensitivity to the negation. $\text{BERT}_{\text{BASE}}$ assigns true statements higher probability than false for 75% of natural sentences ("NT"), and $\text{BERT}_{\text{LARGE}}$ does so for 87.5% of natural sentences. By contrast, the models each show preference for true statements in only 37.5% of items designed to be less natural ("LN"). Table 14 shows these sensitivities broken down by affirmative and negative conditions. Here we see that in the natural sentences, BERT prefers true statements for both affirmative and negative contexts—by contrast, the less natural sentences show the pattern exhibited on NEG-88-SIMP, in which BERT prefers true statements in a high proportion of affirmative sentences, and in 0% of negative sentences, suggesting that once again BERT is defaulting to category matches with the subject.

	Aff. NT	Neg. NT	Aff. LN	Neg. LN
$\text{BERT}_{\text{BASE}}$	62.5	87.5	75.0	0.0
$\text{BERT}_{\text{LARGE}}$	75.0	100	75.0	0.0

Table 14: Percent of NEG-88-NAT with true continuation given higher probability than false. Aff = affirmative, Neg = negative. NT = natural, LN = less natural.

Table 15 contains $\text{BERT}_{\text{LARGE}}$ predictions on two pairs of sentences from the "Natural" sentence set. It is worth noting that even when BERT's first prediction is appropriate in the context, the top candidates often contradict each other (e.g., *difficult* and *easy*). We also see that even with these natural items, sometimes the negation is not enough to reverse the completions, as with the second pair of sentences, in which the fast food dinner both is and isn't a romantic first date.

9 Discussion

Our three diagnostics allow for a clarified picture of the types of information used for predictions by pre-trained BERT models. On CPRAG-34, we see that both models can predict the best completion approximately half the time (at $k = 5$), and that both models rely non-trivially on word order and full sentence context. However, successful predictions in the face of perturbations also suggest that some of BERT's success on these items may exploit certain loopholes, and when we examine predictions on challenging items, we see clear weaknesses in the commonsense and pragmatic inferences targeted by this set. Sensitivity tests show that BERT can also prefer good completions to bad semantically-related completions in a majority of items, but many of these probability differ-

Context	$\text{BERT}_{\text{LARGE}}$ predictions
Most smokers find that quitting is very ____	*difficult, easy, effective, dangerous, hard*
Most smokers find that quitting isn't very ____	*effective, easy, attractive, difficult, succcessful*
A fast food dinner on a first date is very ____	*good, nice, common, romantic, attractive*
A fast food dinner on a first date isn't very ____	*nice, good, romantic, appealing, exciting*

Table 15: NEG-88-NAT predictions by $\text{BERT}_{\text{LARGE}}$

ences are very small, suggesting that the model's sensitivity is much less than that of humans.

On ROLE-88, BERT's accuracy in matching top human predictions is much lower, with $\text{BERT}_{\text{LARGE}}$ at only 37.5% accuracy—and only 53% even on the most constraining contexts. Perturbations reveal interesting model differences, suggesting that $\text{BERT}_{\text{LARGE}}$ has more sensitivity than $\text{BERT}_{\text{BASE}}$ to the interaction between subject and object nouns. Sensitivity tests reveal that both models are able to use noun position to prefer good completions to role reversals, but the differences are on average even smaller than for CPRAG-34, indicating again that model sensitivity to the distinctions is far less than that of humans. The models' general ability to distinguish role reversals suggests that the low word prediction accuracies are not due to insensitivity to word order *per se*, but rather to weaknesses in event knowledge or understanding of semantic role implications.

Finally, NEG-88 allows us to zero in with particular clarity on a divergence between BERT's predictive behavior and what we might expect from a model using all available information about word meaning and truth/falsity. When presented with simple sentences describing category membership, BERT shows a complete inability to prefer true over false completions for negative sentences. The model shows an impressive ability to associate subject nouns with their hypernyms, but when negation reverses the truth of those hypernyms, BERT continues to predict them nonetheless. By contrast, when presented with sentences that are more "natural", BERT does reliably prefer true completions to false, with or without negation. Although these latter sentences are designed to differ in naturalness, in all likelihood it is not naturalness *per se* that drives the model's relative success on them—but rather a higher frequency of these types of statements in the training data.

The latter result in particular serves to highlight a stark, but ultimately unsurprising, observation about what these pre-trained language models bring to the table. While the function of language processing for humans is to compute meaning and make judgments of truth, language models are trained as predictive models—they will simply leverage the most reliable cues in order to optimize their predictive capacity. For a phenomenon like negation, which is often not conducive to clear predictions, such models may not be equipped to learn the implications of this word's meaning.

10 Conclusion

In this paper we have introduced a suite of diagnostic tests for language models, to better our understanding of the linguistic competencies acquired by pre-training via language modeling. We draw our tests from psycholinguistic studies, allowing us to target a range of linguistic capacities by testing word prediction accuracies and sensitivity of model probabilities to linguistic distinctions. As a case study, we apply these tests to analyze strengths and weaknesses of the popular BERT model, finding that it shows sensitivity to role reversal and same-category distinctions, albeit less than humans, and it succeeds with noun hypernyms, but it struggles with challenging inferences and role-based event prediction—and it shows clear failures with the meaning of negation.

The capacities targeted by these test sets are by no means comprehensive, and future work can build on the foundation of these datasets to expand to other aspects of language processing. Because these sets are small, we must also be conservative in the strength of our conclusions—future work can expand to verify the generality of these results. In parallel, we hope that the weaknesses highlighted by these diagnostics can help to identify areas of need for establishing robust and generalizable models for language understanding.

Acknowledgments

We would like to thank Tal Linzen and Kevin Gimpel for helpful comments, as well as other members of the Toyota Technological Institute at

Chicago for useful discussion. We are also grateful to three anonymous reviewers for valuable feedback on an earlier version of this paper.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *International Conference on Learning Representations*.

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Luisa Bentivogli, Raffaella Bernardi, Marco Marelli, Stefano Menini, Marco Baroni, and Roberto Zamparelli. 2016. SICK through the SemEval glasses. Lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Language Resources and Evaluation*, 50(1):95–124.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Wing-Yee Chow, Cybelle Smith, Ellen Lau, and Colin Phillips. 2016. A 'bag-of-arguments' mechanism for initial verb predictions. *Language, Cognition and Neuroscience*, 31(5):577–596.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loic Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2126–2136. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801.

Kara D Federmeier and Marta Kutas. 1999. A rose by any other name: Long-term memory structure and sentence processing. *Journal of memory and Language*, 41(4):469–495.

Ira Fischler, Paul A Bloom, Donald G Childers, Salim E Roucos, and Nathan W Perry Jr. 1983. Brain potentials related to stages of sentence verification. *Psychophysiology*, 20(4):400–409.

SL Frank, LJ Otten, G Galli, and G Vigliocco. 2013. Word surprisal predicts n400 amplitude during reading. In *ACL 2013-51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, volume 2, pages 878–883.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

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: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1195–1205.

Marta Kutas and Steven A Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161.

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. *arXiv preprint arXiv:1808.09031*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Mante S Nieuwland and Gina R Kuperberg. 2008. When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, 19(12):1213–1218.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1525–1534.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2018. What do you learn from context? probing for sentence structure in contextualized word representations. *ICLR*.

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 *Advances in Neural Information Processing Systems*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *EMNLP 2018*, page 353.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.