Grammaticality and Language Modelling

Jingcheng Niu and **Gerald Penn**
Department of Computer Science
University of Toronto
Toronto, Canada
{niu,gpenn}@cs.toronto.edu

Abstract

Ever since Pereira (2000) provided evidence against Chomsky's (1957) conjecture that statistical language modelling is incommensurable with the aims of grammaticality prediction as a research enterprise, a new area of research has emerged that regards statistical language models as "psycholinguistic subjects" and probes their ability to acquire syntactic knowledge. The advent of The Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) has earned a spot on the leaderboard for acceptability judgements, and the polemic between Lau et al. (2017) and Sprouse et al. (2018) has raised fundamental questions about the nature of grammaticality and how acceptability judgements should be elicited. All the while, we are told that neural language models continue to improve.

That is not an easy claim to test at present, however, because there is almost no agreement on how to measure their improvement when it comes to grammaticality and acceptability judgements. The GLUE leaderboard bundles CoLA together with a Matthews correlation coefficient (MCC), although probably because CoLA's seminal publication was using it to compute inter-rater reliabilities. Researchers working in this area have used other accuracy and correlation scores, often driven by a need to reconcile and compare various discrete and continuous variables with each other.

The score that we will advocate for in this paper, the point biserial correlation, in fact compares a discrete variable (for us, acceptability judgements) to a continuous variable (for us, neural language model probabilities). The only previous work in this area to choose the PBC that we are aware of is Sprouse et al. (2018), and that paper actually applied it backwards (with some justification) so that the language model probability was treated as the discrete binary variable by setting a threshold.

With the PBC in mind, we will first reappraise some recent work in syntactically targeted linguistic evaluations (Hu et al., 2020), arguing that while their experimental design sets a new high watermark for this topic, their results may not prove what they have claimed. We then turn to the task-independent assessment of language models as grammaticality classifiers. Prior to the introduction of the GLUE leaderboard, the vast majority of this assessment was essentially anecdotal, and we find the use of the MCC in this regard to be problematic. We conduct several studies with PBCs to compare several popular language models. We also study the effects of several variables such as normalization and data homogeneity on PBC.

1 Background

The three currently most popular means of evaluating a neural language model are: (1) perplexity, an information-theoretic measure that was in use long before neural networks became the preferred means of implementing language models; (2) task performance profiles, in which derivative aspects of a language model's predictions are embedded in a so-called "downstream" task, with all other aspects of the implementation held constant; and (3) targeted linguistic evaluations, the purpose of which is to demonstrate specific syntactic generalizations that a candidate model implicitly captures or does not capture. These targeted evaluations must take place on a large number of small data sets in order to control for the syntactic and lexical variations that we witness among sentences in a realistic corpus.

The purpose of this paper is ultimately to find a task-independent means of testing how well language model probabilities might serve as grammaticality regression scores. Using evidence from targeted linguistic evaluations, we argue for the point-biserial correlation as at least the basis of

Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP), pages 110–119,
November 20, 2020.

such a task-independent measure, and then use the PBC to examine several neural models along with some important variables that affect both their evaluation and the data that we evaluate on.

Borrowing a convention from linguistic theory, Marvin and Linzen (2018) coined the use of "minimal pairs" as input to language models in order to test these fine-grained variations. For example:

(1) Reflexive pronoun in a sentential complement:

  a. The bankers thought the pilot embarrassed himself.

  b. *The bankers thought the pilot embarrassed themselves.

(2) Reflexive pronoun across an object relative clause:

  a. The manager that the architects like doubted himself.

  b. *The manager that the architects like doubted themselves.

These pairs deal with referential agreement in specific syntactic environments. If a model assigns the grammatical string in a pair a higher score than the ungrammatical string, then we say that the model made the correct prediction on that pair. Having evaluated the model over a large number of these pairs, we can compute an accuracy score, relative to a 50% random baseline.

Hu et al. (2020) have taken exception to the design of many such evaluations on that grounds that: (1) a number of English nouns are stereotypically gendered, which conditions pronoun choice, and (2) the unigram probabilities of reflexive pronouns are different, which biases the probabilities that models assign to sentences that contain them. To circumvent these shortcomings, they generalized the pairs to larger sets of strings in which multiple nouns were used in multiple positions so that lexical choice and order could be permuted across sets. They also introduced *distractors*, grammatical strings that contain material irrelevant, or distracting, to the determination of the sentence's grammaticality. One set that they use, for example, is:

(1B) The girl said that the mother saw herself.

(2B) The mother said that the girl saw herself.

(1D) The girls said that the mother saw herself.

(2D) The mothers said that the girl saw herself.

(1U) The girl said that the mothers saw herself.

(2U) The mother said that the girls saw herself.

where (B) is a baseline grammatical string, (D) is a distractor, and (U) is an ungrammatical string. This set has six strings, but sets in their experiments can have as many as 48 strings each, with as many as 75 sets in a single experiment, each one having a unique target pronoun in all of its strings. Because here it is the context that varies, rather than the pronoun, Hu et al. (2020) must rank the conditional probabilities of the pronoun in these various contexts, rather than total sentence probabilities.

Hu et al. (2020) also evaluate models with accuracies. Because there are three classes of string, rather than two, a model is said to have made the correct prediction if the ungrammatical data receive a lower score than both the baseline and distractor data. But because there are more than three strings, they do not compare individual scores from the candidate model, but rather the three means that result from averaging the conditional pronoun probabilities of the baseline, distractor and ungrammatical strings, respectively.

This alternative design not only provided better accuracies than were achieved by Marvin and Linzen (2018), but the inclusion of distractors in the design lowers the random baseline from 50% to 33.3% accuracy. Hu et al. (2020) conclude that current neural language models are learning more about the licensing of reflexive anaphora than was previously thought.

2 Theoretical Exceptions

In a typical psycholinguistics experiment, we would give human subjects a task to perform during which they would be presented with a stimulus that was labelled as either baseline, distractor or ungrammatical. The effect of the stimulus on the task could be measured by time to completion, the number of correct tokens retrieved during a fixed interval of time, etc. Regardless, the task would almost certainly be chosen so that samples of its corresponding measure of success would be normally distributed. So a within-subjects mean of these quantities is entirely justifiable.

The situation is somewhat less clear with the scores that are returned by a neural language model. Ignoring for the moment that Hu et al. (2020) are interested in conditional pronoun probabilities and

not sentence probabilities, the scores are generally not regarded as measures of success on a task *per se* — there is no actual task here, apart from achieving a high rank in the evaluation. Legitimate task performance profiles are defined over separate downstream tasks, such as those in the GLUE leaderboard (Wang et al., 2018). It is rather more difficult to think of downstream tasks that depend on conditional pronoun probabilities, however. Note that for Marvin and Linzen (2018), the ratio of conditional pronoun probabilities of a set of stimuli was the same as the ratio of their total sentence probabilities because the reflexive pronoun is always the last word of the sentence, and the contexts preceding the pronouns are always identical.

Several papers by Lau et al., culminating in Lau et al. (2017), have argued instead that sentence probabilities can justifiably be interpreted as gradient grammaticality scores, rejecting the long-standing assumption in generative linguistics that grammaticality is a binary judgement. It is also possible to regard sentence probabilities as summaries of group behaviour, such as relative frequencies of binary grammaticality judgements across multiple individual participants, with no claim of gradience implied for any single participant. This in turn raises the very old question of whether neural networks in fact have any cognitive plausibility, which has recently started to be debated again (Cichy and Kaiser, 2019). Sample distributions of means converge to a normal distribution even if the underlying population distribution is not normal itself, and so whether an average would be justified in this group interpretation would depend to a great extent on the sizes of the sets of strings (relatively small, as we have seen) as well as how skewed the underlying distribution was.

3 Empirical Exceptions

3.1 Significance Test: Normality

Using Hu et al.'s (2020) publicly available experimental results, [1] we administered Levene's test of homoscedasticity to every set of probabilities, given a fixed stimulus set, model and experimental context. Levene's test attempts to reject the null hypothesis that a set of continuous data is normal. Levene's test was successful for 22.5% of Hu et al.'s (2020) sets at a confidence threshold of $\alpha = 0.05$, and marginally successful for an additional 8% at $\alpha = 0.1$. This means that somewhere between 20–30% of the sets are provably not normal. Homoscedasticity is merely one aspect of normal distributions that can be used to prove that a distribution is not normal.

[1] https://github.com/jennhu/reflexive-anaphor-licensing.

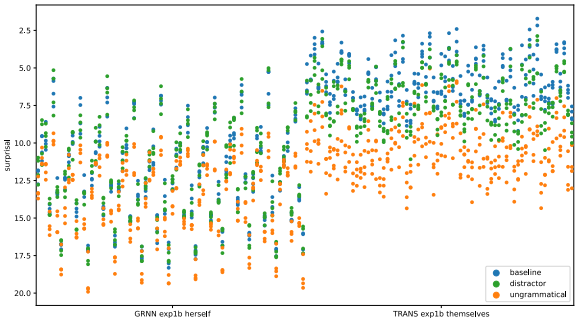

Figure 1: Surprisals (negative log probabilities) for every set in experiment 1b for the GRNN model with *herself*, on the left, and for the TransXL model with *themselves* on the right.

3.2 Significance Test: Mean Differentials

In view of the previous section's results, we elected to use the non-parametric Mann-Whitney U-test to determine, on a set-by-set basis, whether the probability that: "the score of a grammatical (meaning baseline or distractor) string is *greater* than that of an ungrammatical string" is significantly different from the probability that it is *less*. This does not determine the difference between means because it cannot quantify effect size, nor does it even determine the sign of the difference. This is an alternative, very minimalist way of formalizing that a language model has made the correct prediction — it can simply distinguish grammatical from ungrammatical, somehow.

Let us consider part of Hu et al.'s (2020) experiment 1b as an example, shown in Figure 1. There would be little disagreement that the model on the right (Transformer-XL with the pronoun *themselves*) had made better predictions than the model on the left (an LSTM with the pronoun *herself*), and yet under both of these conditions the accuracy is 100%. Large differences involving strings at either extreme help to offset a number of smaller differences of the wrong sign when computing differences in means.

Across all experiments, 44.3% of the sets in which the mean differentials qualified for the numerator of the accuracy computation (i.e., the ungrammatical mean was less than both the baseline

and distractor means) failed to show a significant difference under the criterion of the Mann-Whitney test. A further 90% of the sets in which the mean differential did not qualify for the numerator (i.e., they were taken not to have been correctly predicted) also failed to show a significant difference. Of the 60 combinations of pronoun and experimental context that we examined, 24 did not have even a single set that showed significance in the numerator. Of the 42 combinations that did not have 100% accuracies, 32 did not have even a single set that showed significance.

In our view, although we agree with every one of the design modifications made by Hu et al. (2020) to targeted evaluations such as these, the decision to continue using accuracy and to generalize it in this way seems not to be working well.

3.3 Matthews Correlation Coefficients

This is potentially a much more pervasive problem than just with Hu et al.'s (2020) experiments. MCCs have emerged as a popular alternative among language modelling enthusiasts (Liu et al., 2019; Lan et al., 2020; Raffel et al., 2019) since grammaticality classification with The Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) was incorporated into the GLUE standard (Wang et al., 2018). Warstadt et al. (2019) themselves began using MCCs, initially to validate the CoLA corpus, but also to interpret Lau et al.'s (2017) gradient models. MCCs cannot be computed directly on continuous data, which means not only that they are insensitive to the magnitudes of probabilities, but also that a threshold must be set in order to impose a discrete boundary between classes. Defending that choice of boundary can be difficult. Consider Figure 2, for example. In a sample as small as a typical minimal set, cross-validating the MCC decision threshold is not realistic, so here we used the mean of both classes of data. In this particular set, two low-surprisal distractors cause a lot of damage to the distractor vs. ungrammatical MCC and the baseline-plus-distractor vs. ungrammatical MCC. Another correlation score, called the point-biserial correlation, which can be computed directly on continuous data, does not require an arbitrary threshold and produces very different values on this one example.

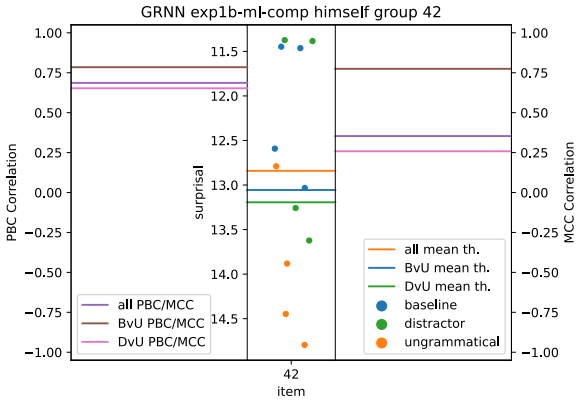

Figure 2: Surprisals (negative log probabilities), Matthews correlations and point-biserial correlations for set 42 in Hu et al.'s (2020) experiment 1b for the GRNN model with the pronoun, *himself*.

4 Aggregated Point Biserial Correlations

Our proposed alternative involves two changes. First, we propose using a point biserial correlation between the output probability of a language model and binary grammaticality judgements. Second, we propose calculating PBCs not on a set-by-set basis, but for all probabilities generated by a fixed model using all of the contexts of a fixed experiment.

To consider Figure 1 again, the model on the left has a PBC of 0.25, whereas the model on the right has one of 0.73. Correlations such as the PBC range between -1 and 1, where 1 is perfect correlation, 0 is no correlation, and -1 is perfect anti-correlation.

Our choice of PBC is perhaps the less controversial of these two changes, as our motivation for doing so is mainly due to the fact that it is *the* standard measure for correlating a continuous or interval random variable with a discrete random variable.

Our decision to "aggregate" data by ignoring the boundaries between the controlled, minimal sets that have become so widely accepted as a part of targeted syntactic evaluations is perhaps counterintuitive. But as long as the necessary distractors, permutations and lexical alternations that avoid bias appear somewhere in the context of the experiment, they will be compared to each other, although along with additional comparisons that were not made when accuracy was averaged over sets. Those additional comparisons, however, will merely corroborate the model's (in)ability to more robustly distinguish between well-formed and non-

well-formed strings, and the experiment itself does restrict the variability of those comparisons to a great extent.

In our experience, aggregating makes the evaluation more resilient to choices of normalizers such as SLOR (Pauls and Klein, 2012), its results are in closer accord to our intuitive judgements, and, as expected, it handles sample bias better. Both accuracy (30–100%) and aggregate PBC (-0.01–0.81) vary widely from experiment to experiment in Hu et al.'s (2020) data, and yet the average of per-set PBCs tends to be less dispersed. The experiments in Figure 1, for example, have microaveraged PBCs of 0.77 (left) and 0.89 (right). It could therefore be argued that the effect size of the dependent variable that Hu et al. (2020) were attempting to measure is not as large as the choice of minimal set. Aggregation would then also be an effective means of utilizing the available range of correlation values.

Total (weighted) accuracy and baseline-plus-distractor vs. ungrammatical PBC have a Spearman's correlation of 0.876 ($p = 8.5 \times 10^{-25}$) across Hu et al.'s (2020) experiments and models.

5 Task-Independent Grammaticality Classification

The famous "Colorless green ideas sleep furiously" (CGISF) example (Chomsky, 1957) posited a seemingly irreconcilable divide between formal linguistics and statistical language modelling, arguing that every sequence of words not attested in the collective memory of a language's use would be considered equally "remote" by a putative instance of the latter, regardless of whether the sequence was grammatical (CGISF) or ungrammatical. The example was presented briefly and informally in order to reject statistical language modelling as an alternative approach to the one advocated and developed in greater detail by Chomsky (1957). It was only presented with one other example, the reverse of the sentence, i.e., "Furiously sleep ideas green colorless", in order to draw a contrast between two nonsensical sequences, only one of which (CGISF) is grammatical.

Pereira (2000) provides an attempt at a refutation by constructing a statistical language model based upon an agglomerative Markov process (Saul and Pereira, 1997), and then observing that CGISF is assigned a probability by the model which is roughly 200 000 times greater than the probability assigned to its reversal.

There has nevertheless been some scepticism expressed about the ensuing euphoria among computer scientists — mainly by linguists. Sprouse et al. (2015) notes that the trigram model from Lau et al. (2015) assigns different rankings to 10 different permutations of CGISF, depending on the training corpus (e.g., the *Wall Street Journal* corpus versus an example training corpus taken from Lau et al. (2015)). Can the scores assigned to these sequences be reliably construed as a regression scale of grammaticality (or perhaps acceptability), if they are so fickle? Chowdhury and Zamparelli (2018) also express concern about the ability of neural language models to generalize to more abstract grammatical phenomena than subject-verb agreement.

What we will present in this section is a more thorough appraisal of how well statistical language models perform as instruments of grammaticality testing overall, using PBCs. Previous research on grammaticality/acceptability and language models has mainly designed experiments using naturally occurring English sentences, and modifies those sentences based on various individual linguistic phenomena to manually introduce a specific source of ungrammaticality into the sentences. Notable exceptions include CoLA as well as the Linguistic Inquiry (LI) corpus of grammatical and ungrammatical sentences collected by Sprouse et al. (2013) and Sprouse and Almeida (2012), and used in Sprouse et al. (2018). Both are based upon examples found in linguistics publications. Lau et al. (2014, 2015, 2017) create ungrammatical sentences by round-trip translating natural English sentences. We will use both CoLA and the LI corpus.

5.1 CoLA

The Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) is a collection of 10 657 example sentences from linguistics publications with their grammaticality judgements. It forms an integral part of the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018). It must be noted, however, that their linguistic acceptability task is supervised (CoLA was divided into a training set (8551), a development set (1043), and a test set (1063)), with both positive and negative samples. The ungrammatical strings in CoLA have generally been devised to illustrate a specific grammatical defect, and are often but not always sensical. Recent systems trained on these labelled

data, e.g., Liu et al. (2019); Lan et al. (2020); Raffel et al. (2019), are able to attain a reported roughly 70% Matthews correlation coefficient (Matthews, 1975).

The performance of Mikolov's (2012) model, in particular, has been reported in CoLA studies as a baseline (Warstadt et al., 2019; Lau et al., 2017). Warstadt et al. (2019) did a 10-fold cross-validation on CoLA test set, which fit an optimum decision threshold to the softmax output of each fold to assign grammaticality labels, and obtained a 0.652 in-domain accuracy and 0.711 out-of-domain accuracy. This figure has been cited as a gauge for assessing the ability of statistical language models to learn grammar-related patterns in an unsupervised fashion (Lappin and Lau, 2018).

CoLA also did not include any annotations of minimal set structures, but we retained a linguist who is a native speaker of North American English to go over the first 2010 sentences in the CoLA corpus and group them into 1803 microgroups (including singletons) fashioned around the same linguistic phenomena of interest, and often very similar lexical entries. This enabled us to use CoLA as a platform to test language model performance in somewhat controlled microgroups of example sentences, although they are not as well controlled as the minimal sets of targeted evaluations. Then we ran point-biserial correlation tests within those microgroups with at least one grammatical judgement and at least one ungrammatical judgement, and calculated the median of those correlation scores. Then we split the scores into four quadrants. Below, we report the junction points of those quadrants: lower breakpoint, median, and upper breakpoint.

5.2 The LI Corpus

The LI corpus was collected by Sprouse and Almeida (2012), and contains 300 sentence structures, each expanded into 8 candidate sentences (2400 strings in total, 1192 of them grammatical). The corpus annotation shows that there are 57 pairs of sentence structures (912 strings in total) that are syntactically designed to differ on one linguistic phenomenon but have putatively opposite grammaticality. We ran the PBC test for each of the 57 pairs, and calculated the medians among the correlation scores. Sprouse and Almeida (2012) also collected 230 sentences structures from Adger's (2003) textbook. However that corpus does not include annotation indicating the minimal set structure, and therefore was ignored in this study.

5.3 Language Models

We investigated four different types of language models: Pereira's (2000) original aggregative Markov model (Saul and Pereira, 1997), Mikolov's (2012) original RNN language model (Mikolov, 2012), a QRNN-based language model (Merity et al., 2018) that we take to be representative of contemporary models, and GPT-2 (Radford et al., 2019) as the representative of large-scale pre-trained language models. Mikolov's model is also used by Clark et al. (2013); Lau et al. (2015); Sprouse et al. (2018) in their research about gradient acceptability. We chose GPT-2 over other large-scale pre-trained models such as BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019), because it took the more orthodox autoregressive language modelling approach that is consistent with the remaining choices, and it is most commonly used for natural language generation for the same reason.

We obtained a publicly available implementation of each of the four language models[2]. The implementation of the tied adaptive softmax (TAS) method[3] used the unusual approach of applying softmax on already softmaxed values. For this reason, we also experiment on QRNN models trained using regular cross-entropy loss functions.

All three non-pretrained models are trained on the BNC (BNC Consortium, 2007) and WikiText-103 (WT103) (Merity et al., 2017). We used the hyperparameters described in (Pereira, 2000) to train its model, the hyperparameters described in (Lau et al., 2017; Sprouse et al., 2018) to train Mikolov's, and the hyperparameters suggested by the official SalesForce implementation of the QRNN model. The BNC corpus is tokenized based on BNC annotations, and all tokens are converted into lower case. For WT103, we used the official preprocessed corpus released on SalesForce's website[4] that has tokenization, converts low frequency words to *unk* and preserves letter case. Radford et al. (2019) re-

[2]For Pereira's model, we adapted the implementation of https://github.com/hsgodhia/agm_language_model; for Mikolov's model, we used the implementation of https://github.com/yandex/faster-rnnlm that is also used by Lau et al. (2017) and Sprouse et al. (2018); and for GPT-2, we used the HuggingFace's Transformers package (Wolf et al., 2020).

[3]https://github.com/salesforce/awd-lstm-lm

[4]https://blog.einstein.ai/the-wikitext-long-term-dependency-language-modeling-dataset/

Model	BNC	WT103
Pereira	362.19	460.62
Mikolov	332.04	185.82
QRNN (regular)	173.57	96.65
QRNN (TAS)	92.85	34.98
GPT-2[5]	45.61	28.34
GPT-2 XL[5]	24.92	16.28

Table 1: Perplexity achieved on test sets

leased GPT-2 models in four different parameter sizes: GPT2 (small), GPT2-medium, GPT2-large and GPT2-xl (extra large). To avoid redundancy, we experimented with GPT2, which has a similar number of parameters as the other neural language models, and GPT2-xl, which represents the maximum potential of the GPT-2 architecture. Better performance would likely be achieved through more extensive hyperparameter optimization, but our results in Table 1 are already comparable to the performance reported in the respective original publications.

5.4 Experimental Design

Our experiments consider two types of probabilities: the log probability $\ell = \log p(s)$, and the actual probability, e^{ℓ}, where s is a sentence. For each type of probability, we also consider two to three different types of normalization methods: no normalization (raw), normalization by length (norm) $\ell/|s|$, and SLOR (Pauls and Klein, 2012) $(\ell - \ell_u)/|s|$, where $|s|$ is the length of the sentence and ℓ_u is the log unigram probability of the sentence. For all three non-pretrained models, the unigram probability was obtained from BNC/WT103 with add-one smoothing. We used WT103 unigram probabilities for GPT-2 models since they preserve case.

5.5 Letter Case

It is a paradigm that linguists often consider semantics and pragmatics when trying to generate non-syntactic factors that attribute to language model probabilities. We also considered letter case in order to demonstrate that a more superficial fact about the writing system may affect the evaluation result. Pereira's (2000) model downcased all input tokens to speed up the training process, thus it was discarded for this experiment. We took the rest of the models that are trained on WT103 and GPT-2 models and provided them with downcased CoLA example sentences.

[5]GPT-2 models are evaluated on the same preprocessed BNC and WT103 test sets without any fine-tuning for the sake of consistency.

Model	Norm.	BNC		WT103	
		LOG	EXP	LOG	EXP
Pereira	Raw	0.0239	0.0139	0.0226	-0.0137
	Norm	0.0494	0.0206	0.043	-0.0012
	SLOR	0.0756	-0.0153	0.0684	0.0083
Mikolov	Raw	0.0578	0.0223	0.0574	0.0086
	Norm	0.1061	0.1161	0.106	0.1146
	SLOR	0.1896	0.1529	0.1045	0.0359
QRNN regular	Raw	0.0029	-0.0153	0.0121	-0.0093
	Norm	0.0191	0.0328	0.0124	0.0224
	SLOR	0.0496	0.0346	0.0297	0.0134
QRNN TAS	Raw	0.0067	-0.0137	0.0162	0.0047
	Norm	0.0029	-0.0153	0.0278	0.0421
	SLOR	0.0542	0.0356	0.0332	0.0112
		GPT-2		GPT-2 XL	
GPT-2 Models	Raw	0.1839	0.0117	0.1476	0.0123
	Norm	0.2498	0.1643	0.2241	0.1592
	SLOR	0.2489	0.092	0.2729	0.0872

Table 2: CoLA Point-biserial Test Results

Model	Norm.	BNC		WT103	
		LOG	EXP	LOG	EXP
Pereira	Raw	0.0231	-0.0136	-0.0027	-0.0488
	Norm	0.0758	0.0595	0.038	0.0412
Mikolov	Raw	0.0841	0.1868	0.1278	0.1914
	Norm	0.2541	0.2465	0.1955	0.2043
QRNN TAS	Raw	-0.0066	-0.2197	0.0201	0.0186
	Norm	0.068	0.0726	0.1058	0.0764
QRNN regular	Raw	-0.0135	-0.059	0.0232	0.1251
	Norm	0.042	0.0245	0.1057	0.104
		GPT-2		GPT-2 XL	
GPT-2 Models	Raw	0.4671	0.26	0.4767	0.266
	Norm	0.5233	0.487	0.5653	0.5131

Table 3: Sprouse LI Minimal Sets Results

5.6 "Sensicality"

Can we find anything that matches language model outputs better than a grammaticality judgement? Inspired by the debate over "Colorless green ideas sleep furiously" sixty years ago, we formed the hypothesis that grammatical sentences that make sense could more easily be distinguished from grammatical sentences that are nonsense. We formulated 27 nonsense sentences (including CGISF), projected their parts of speech into the BNC and found 36 exact POS matches that do not overlap with a clause or sentence boundary. The "sensicality" task is to distinguish these two sets using language model log-probabilities.

6 Experiment Results

CoLA Point-Biserial Correlation Test Table 2 shows our PBC test results. As mentioned before, every non-GPT-2-based model is trained on either BNC or WT103, and for the sake of simplicity, we report two sizes of GPT-2: small and XL.

All models show weak to no correlation. However the correlation generated by GPT-2 models does show significantly greater promise.

LI Minimal Sets Table 3 shows the language models' performance on the LI minimal sets.

Model	Norm.	Score	all group sizes (≥ 2)			group size > 4		
			median	up bkpt.	low bkpt.	median	up bkpt.	low bkpt.
Pereira BNC	raw	log	0.2148	0.8119	-0.4193	0.1409	0.4161	-0.0908
	raw	exp	0.3233	0.641	-0.4616	0.2501	0.3598	-0.2048
	Norm	log	0.3787	0.9026	-0.309	0.1865	0.4706	-0.1122
	Norm	exp	0.3573	0.867	-0.3238	0.2753	0.4537	-0.1087
Pereira WT103	raw	log	0.2383	0.8278	-0.3901	0.159	0.4084	-0.1335
	raw	exp	0.3254	0.6642	-0.4874	0.2244	0.3466	-0.1778
	Norm	log	0.3601	0.8849	-0.2933	0.2381	0.4849	-0.1449
	Norm	exp	0.3599	0.9023	-0.299	0.2108	0.4438	-0.0941
Mikolov BNC	raw	log	0.3838	0.8383	-0.294	0.2462	0.5453	-0.1127
	raw	exp	0.3834	0.8092	-0.3073	0.262	0.4274	-0.1617
	Norm	log	0.291	0.8411	-0.4262	0.2159	0.4823	-0.2066
	Norm	exp	0.3314	0.6903	-0.5	0.2382	0.4086	-0.0918
Mikolov WT103	raw	log	0.3651	0.8506	-0.3714	0.2701	0.4835	-0.0786
	raw	exp	0.3516	0.7988	-0.4765	0.2577	0.4199	-0.1876
	Norm	log	0.4986	0.9303	-0.1553	0.2996	0.5332	-0.0953
	Norm	exp	0.4918	0.9363	-0.1417	0.3012	0.5495	-0.0538
QRNN regular BNC	raw	log	0.0567	0.6549	-0.5812	0.0602	0.3224	-0.2602
	raw	exp	0.2	0.5	-0.5633	0.1961	0.3085	-0.2605
	Norm	log	0.0418	0.682	-0.4762	-0.0436	0.3349	-0.3539
	Norm	exp	0.0308	0.6565	-0.5317	-0.0275	0.3296	-0.3205
QRNN regular WT	raw	log	0.0249	0.6086	-0.6074	0.0534	0.4419	-0.2144
	raw	exp	0.2112	0.51	-0.5337	0.2483	0.3728	-0.1608
	Norm	log	0.0507	0.8203	-0.5342	0.0291	0.3456	-0.3227
	Norm	exp	0.084	0.8547	-0.5374	0.0428	0.3743	-0.2727
QRNN TAS BNC	raw	log	0.0456	0.5834	-0.6022	0.1171	0.4084	-0.2532
	raw	exp	0.1487	0.5	-0.5595	0.2003	0.3268	-0.2474
	Norm	log	0.0919	0.6522	-0.5106	-0.0073	0.3629	-0.3003
	Norm	exp	0.1248	0.6203	-0.5312	0.0379	0.3507	-0.3004
QRNN TAS WT103	raw	log	0.1144	0.7072	-0.622	0.1187	0.4071	-0.1658
	raw	exp	0.2524	0.5003	-0.5773	0.2212	0.3173	-0.2568
	Norm	log	0.2061	0.8411	-0.4252	0.0698	0.3924	-0.2324
	Norm	exp	0.2226	0.8138	-0.4726	0.1536	0.3975	-0.1804
GPT-2	raw	log	0.6256	0.9491	0.1424	0.4128	0.6692	0.1424
	raw	exp	0.4902	0.9999	0.2	0.2823	0.4417	0.1794
	Norm	log	0.7121	0.9914	0.0528	0.2948	0.6688	0.0259
	Norm	exp	0.6597	0.9968	0.1522	0.3294	0.6212	0.1105
GPT-2 XL	raw	log	0.6936	0.9865	0.2862	0.4503	0.7155	0.1953
	raw	exp	0.5	1.0	0.2642	0.2956	0.4714	0.2117
	Norm	log	0.6858	0.9983	0.2411	0.4537	0.6561	0.1803
	Norm	exp	0.6516	0.9987	0.2939	0.4312	0.5988	0.2477

Table 4: CoLA Microgrouping Results

Model	Norm.	LOG		EXP	
		with case	lower	with case	lower
Mikolov	Raw	0.0578	0.0574	0.0223	0.0206
	Norm	0.1061	0.0955	0.1161	0.1012
QRNN regular	Raw	0.0029	0.0086	-0.0153	-0.0186
	Norm	0.0191	0.0135	0.0328	0.0149
QRNN TAS	Raw	0.0067	0.0148	-0.0137	-0.0133
	Norm	0.0309	0.0357	0.0301	0.0146
GPT-2	Raw	0.1476	0.1129	0.0123	0.0125
	Norm	0.2241	0.1968	0.1592	0.1403
GPT-2 XL	Raw	0.1839	0.1484	0.0117	0.0149
	Norm	0.2498	0.2057	0.1643	0.1372

Table 5: Letter Case Study Results

Again, the GPT-2 models stand out, but in this case, GPT2-xl performs consistently better.

CoLA Microgroups Table 4 shows the microgrouping results. The results could be interpreted as confirming our hypothesis: that better controlled input would improve a language model's ability to focus on distinguishing grammaticality. On the other hand, it is also likely that the very small size of most microgroups is a factor, because there is a dramatic correlation drop when we evaluate on microgroups with size greater than 4. Roughly 77% of the non-singleton microgroups in CoLA are of size 2-4.

Letter Case Table 5 shows the letter case study's result. GPT-2 is once again the best, but it also suffers the most from the loss of case. The loss is comparable to the loss incurred by scaling the XL model's size (1542M) back to small (117M).

Model	Norm.	BNC	WT	GPT-2
Pereira	raw	0.8235	0.7652	
	SLOR	0.1838	0.1927	
Mikolov	raw	0.827	0.9042	
	SLOR	-0.3161	-0.1556	
QRNN	raw	0.7132	-0.3872	
	SLOR	0.089	-0.8038	
QRNN-R	raw	0.8064	0.5895	
	SLOR	0.6598	-0.7192	
GPT-2	raw			0.7574
	SLOR			0.5486
GPT-2 XL	raw			0.7642
	SLOR			0.5218

Table 6: Sensicality Results

Sensicality The sensicality study reveals much higher PBC scores overall, although SLOR has a markedly detrimental effect overall. While this set of judgements is small, these scores are markedly higher than the PBCs for the microgroupings as well, all but one of which is smaller.

7 Discussion

In this paper, we examined the motivation and effects of using accuracy scores vs. PBC in syntactically targeted models. We also used PBC to evaluate a range of language models on curated datasets. While the results are not terribly strong, GPT-2's showing in particular suggests that a great deal of progress has been made recently.

It is nevertheless still premature to claim that the probabilities assigned by language models to sequences of words can be reliably construed as a regression scale of grammaticality. Such a claim would need to be supported by a stronger performance in more diverse settings that are larger than minimal-set or microgrouping structures, ideally with better robustness to other factors such as type case. The sensicality study suggests that language models are still overwhelmingly influenced by semantic factors. This is unsurprising: language models have been used for years as a proxy for semantics in numerous other areas such as parsing.

The best grammaticality classifiers to date are still classifiers that are constructed for the purpose of predicting grammaticality, not for the classical purpose of a language model, which is to predict the next word of input. These either use a language model output probability as their own input (Warstadt et al., 2019) or use other artefacts of the language model, such as word vectors, and discard the language model probability altogether (Liu et al., 2019).

Acknowledgments

We would like to thank Zoe McKenzie for her grammaticality judgements.

References

David Adger. 2003. *Core Syntax: A Minimalist Approach*, volume 20.

BNC Consortium. 2007. The British National Corpus, version 3 (BNC XML Edition). *Distributed by Oxford University Computing Services on behalf of the BNC Consortium.*

Noam Chomsky. 1957. *Syntactic structures*. Mouton publishers.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN Simulations of Grammaticality Judgments on Long-distance Dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Radoslaw M. Cichy and Daniel Kaiser. 2019. Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, 23(4):305–317.

Alexander Clark, Gianluca Giorgolo, and Shalom Lappin. 2013. Statistical Representation of Grammaticality Judgements: The Limits of N-Gram Models. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 28–36, Sofia, Bulgaria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*.

Jennifer Hu, Sherry Chen, and Roger Levy. 2020. 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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Shalom Lappin and Jey Han Lau. 2018. Gradient Probabilistic Models vs Categorical Grammars: A Reply to Sprouse et al.(2018).

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2014. Measuring Gradience in Speakers' Grammaticality Judgements. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36:6.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised Prediction of Acceptability Judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1618–1628, Beijing, China. Association for Computational Linguistics.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

B. W. Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. An Analysis of Neural Language Modeling at Multiple Scales. *arXiv:1803.08240 [cs]*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer Sentinel Mixture Models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Tomáš Mikolov. 2012. Statistical language models based on neural networks. *Brno University of Technology dissertation*.

Adam Pauls and Dan Klein. 2012. Large-Scale Syntactic Language Modeling with Treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968, Jeju Island, Korea. Association for Computational Linguistics.

Fernando Pereira. 2000. Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 358(1769):1239–1253.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv e-prints*.

Lawrence Saul and Fernando Pereira. 1997. Aggregate and mixed-order Markov models for statistical language processing. In *Second Conference on Empirical Methods in Natural Language Processing*.

Jon Sprouse and Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax1. *Journal of Linguistics*, 48(3):609–652.

Jon Sprouse, Sagar Indurkhya, Sandiway Fong, and Robert C. Berwick. 2015. Colorless green ideas do sleep furiously: The necessity of grammar. *The 46th Annual Meeting of North East Linguistic Society (NELS 46)*.

Jon Sprouse, Carson T. Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, 134:219–248.

Jon Sprouse, Beracah Yankama, Sagar Indurkhya, Sandiway Fong, and Robert C. Berwick. 2018. Colorless green ideas do sleep furiously: Gradient acceptability and the nature of the grammar. *The Linguistic Review*, 35(3):575–599.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural Network Acceptability Judgments. *arXiv:1805.12471 [cs]*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.