Neural Network Acceptability Judgments

Alex Warstadt[1]
warstadt@nyu.edu

Amanpreet Singh[2]
amanpreet@nyu.edu

Samuel R. Bowman[1,2,3]
bowman@nyu.edu

[1]Dept. of Linguistics
New York University
10 Washington Place
New York, NY 10003

[2]Dept. of Computer Science
New York University
60 Fifth Avenue
New York, NY 10011

[3]Center for Data Science
New York University
60 Fifth Avenue
New York, NY 10011

Abstract

This paper investigates the ability of artificial neural networks to judge the grammatical acceptability of a sentence. Machine learning research of this kind is well placed to answer important open questions about the role of prior linguistic bias in language acquisition by providing a test for the Poverty of the Stimulus Argument. In service of this goal, we introduce the Corpus of Linguistic Acceptability (CoLA), a set of 10,657 English sentences labeled as grammatical or ungrammatical from published linguistics literature. As baselines, we train several recurrent neural network models for acceptability classification. These models show promise on the task, and error-analysis on specific grammatical phenomena reveals that they learn some systematic generalizations like subject-verb-object word order without any grammatical supervision. However, human-like performance across a wide range of grammatical constructions remains far off.

1 Introduction

Native English speakers consistently report a sharp contrast in acceptability[1] between pairs of sentences like (1), irrespective of their grammatical training.

(1) a. What did Betsy paint a picture of?
b. *What was a picture of painted by Betsy?

Such *acceptability judgments* are the primary source of empirical data in much of theoretical linguistics, and the goal of a *generative grammar* is to generate all and only those sentences which native speakers find acceptable (Chomsky, 1957; Schütze, 1996).

Despite this centrality, there has been relatively little work on acceptability classification in computational linguistics. The recent explosion of progress in deep learning inspires us to revisit this task. The task has important implications for theoretical linguistics as a test of the Poverty of the Stimulus Argument, as well as in natural language processing as a way to probe the grammatical knowledge of neural sequence models.

The primary contribution of this paper is to introduce a new dataset and several novel neural network baselines with the aim of facilitating machine learning research on acceptability. To begin we define and motivate a version of the acceptability classification task that is suitable for sentence-level machine learning experiments (Section 2). We address the lack of readily available acceptability judgment data by introducing the Corpus of Linguistic Acceptability (CoLA), a collection of sentences labeled for acceptability from the linguistics literature, which at 10,657 examples is by far the largest of its kind.

We train several semi-supervised neural sequence models to do acceptability classification on CoLA and compare their performance with unsupervised models from Lau et al. (2016). Our best model outperforms unsupervised baselines, but falls short of human performance on CoLA by a wide margin. We conduct an error analysis to test our models' performance on specific linguistic phenomena,

[1]Following conventions in linguistics, *acceptability* in this paper is a notion of linguistic performance which can be observed by introspective judgments, while *grammaticality* is an abstract notion of linguistic competence (Schütze, 1996).

Included	Morphological Violation	(a)	*Maryann should leaving.
	Syntactic Violation	(b)	*What did Bill buy potatoes and _?
	Semantic Violation	(c)	*Kim persuaded it to rain.
Excluded	Pragmatical Anomalies	(d)	*Bill fell off the ladder in an hour.
	Unavailable Meanings	(e)	*He$_i$ loves John$_i$. (*intended*: John loves himself.)
	Prescriptive Rules	(f)	Prepositions are good to end sentences with.
	Nonce Words	(g)	*This train is arrivable.

Table 1: Our classification of unnacceptable sentences, shown with their presence or absence in CoLA.

and find that some models systematically distinguish certain kinds of minimal pairs of sentences differing in gross word order and argument structure. Our experiments show that recurrent neural networks can beat strong baselines on the acceptability classification task, but there remains considerable room for improvement.

1.1 Resources

CoLA can be downloaded from the CoLA website.[2] The site also hosts a demo of our best model. Our code is available as well.[3] There are also two competition sites for evaluating acceptability classifiers on CoLA's in-domain[4] and out-of-domain[5] test sets.

2 Acceptability Judgments

2.1 In Linguistics

Acceptability judgments are central to the formulation of generative linguistics in Chomsky's influential (1957) book *Syntactic Structures*:

> The fundamental aim in the linguistic analysis of a language L is to separate the grammatical sequences which are the sentences of L from the ungrammatical sequences which are not sentences of L and to study the structure of the grammatical sequences. [...] One way to test the adequacy of a grammar proposed for L is to determine whether or not the sequences that it generates are actually grammatical, i.e., acceptable to a native speaker. (p.13)

This has been the predominant methodology for research in generative linguistics over the last sixty years (Chomsky, 1957; Schütze, 1996). Linguists generally provide example sentences annotated with binary acceptability judgments from themselves or several native speakers.

2.2 The Acceptability Classification Task

Following common practice in linguistics, we define acceptability classification as a binary classification task. An acceptability classifier, then, is a function that maps strings into the set $\{0, 1\}$, where '0' is interpreted as unacceptable and '1' as acceptable. This definition also includes generative grammars of the type described by Chomsky (1957) above.

CoLA consists entirely of examples from the linguistics literature. Linguists generally present examples to motivate specific arguments, and these sentences are generally chosen to each isolate a particular grammatical construction while minimizing potential distractions. In other words, ungrammatical examples in linguistics publications, like those in CoLA, tend to be unacceptable for a single identifiable reason.

2.3 Defining (Un)acceptability

Not all linguistics examples are suitable for acceptability classification. While all acceptable sentences can be included, we exclude four types of unacceptable sentences from the task (examples in Table 1):

Pragmatic Anomalies Examples like (d) can be made interpretable, but only in fanciful scenarios, the construction of which requires real-world knowledge unrelated to grammar.

[2] https://nyu-mll.github.io/CoLA/

[3] https://github.com/nyu-mll/CoLA-baselines

[4] https://www.kaggle.com/c/cola-in-domain-open-evaluation

[5] https://www.kaggle.com/c/cola-out-of-domain-open-evaluation

Unavailable Meanings Examples like (e) are often used to illustrate that a sentence cannot express a particular meaning. This example can only express that someone other than John loves John. We exclude these examples because there is no simple way to force an acceptability classifier to consider only the interpretation in question.

Prescriptive Rules Examples like (f) violate rules which are generally explicitly taught rather than being learned naturally, and are therefore not considered a part of native speaker grammatical knowledge in linguistic theory.

Nonce Words Examples like (g) illustrate impossible affixation or lexical gaps. Since these words will not appear in the vocabularies of typical word-level NLP models, they will be impossible for these models to judge.

The acceptability judgment task as we define it still requires identifying challenging grammatical contrasts. A successful model needs to recognize (a) morphological anomalies such as mismatches in verbal inflection, (b) syntactic anomalies such as wh-movement out of extraction islands, and (c) semantic anomalies such as violations of animacy requirements of verbal arguments.

2.4 Concerns about Acceptability Judgments

Binary vs. Gradient Judgments Though discrete *binary* acceptability judgments are standard in generative linguistics (Schütze, 1996), Lau et al. (2016) find that when speakers are presented with the option to use a gradient scale to report sentence acceptability, they predictably and systematically use the full scale, rather than clustering their judgments near the extremes as would be expected for a fundamentally binary phenomenon. This is evidence, they argue, that acceptability judgments are gradient in nature. Nevertheless, we consider binary judgments in published examples sufficient for our purposes. These examples are generally chosen to be unambiguously acceptable or unacceptable, and provide the evidence that relevant experts consider maximally germane the questions at hand.

Reliability of Judgments Gibson and Fedorenko (2010) express concern about standard practices around acceptability judgments and call for theoretical linguists to quantitatively measure the reliability of the judgments they report, sparking an ongoing dialog about the validity and reproducibility of these judgments (Sprouse and Almeida, 2012; Sprouse et al., 2013). We take no position on this general question, but perform a small human evaluation to gauge the reproducibility of the judgments in CoLA (Section 5).

3 Motivation

The acceptability judgment task offers a direct way to test the *Poverty of the Stimulus Argument*, a key argument in the theory of a strong Universal Grammar (Chomsky, 1965). In addition, acceptability classifiers can be used to probe the grammatical knowledge of neural network models by enabling researchers to test their models for knowledge of specific grammatical constructions.

3.1 The Poverty of the Stimulus

The Poverty of the Stimulus Argument holds that purely data-driven learning is not powerful enough to explain the richness and uniformity of human grammars, particularly with data of such low quality as children are exposed to (Clark and Lappin, 2011). This argument is generally wielded in support of the theory of a strong Universal Grammar, which claims that all humans share an innately-given set of language universals, and that domain-general learning procedures are not sufficient to acquire language (Chomsky, 1965).

There is a need to test whether data-driven learners are indeed able to do acceptability classification within constraints similar to those of human learners. For these experiments to have any bearing on this argument, the artificial learner must not be exposed to any knowledge of language that could not plausibly be part of the *input* to a human learner. We call such knowledge *grammatical bias*. For example, training a learner with a part-of-speech tagging objective or with extremely large datasets would expose the model to rich linguistic knowledge far beyond what human learners see. Our experiments as described in Section 6 are designed to meet these standards (see Section 6.2 for discussion).

3.2 Investigating the Black Box

Recurrent neural network models like the Long Short-Term Memory (LSTM) networks we use (Hochreiter and Schmidhuber, 1997) can discover some forms of structure in raw language data (LeCun et al., 2015). These models are widely used to encode features of sentences in fixed-length sentence embeddings (Cho et al., 2014; Sutskever et al., 2014; Kiros et al., 2015). Evaluating general-purpose sentence embeddings is an active research area. Some approaches probe the contents of sentence embeddings using common natural language processing tasks (Conneau et al., 2017; Wang et al., 2018). Others test whether embeddings encode top level features like sentence length, parse-tree depth, and tense (Adi et al., 2016; Shi et al., 2016; Conneau et al., 2018).

Acceptability classification can be used to probe sentence representations for linguistic features at a finer level of granularity. After training an acceptability classifier with the sentence representation as its input, it is simple to ask how well that representation encodes linguistic phenomena like thematic role, animacy, and anaphoric dependency simply by testing the classifier on appropriate examples. Section 8 discusses several such case studies.

Linzen et al. (2016) present a special case of this approach. They train LSTM models to identify violations in a specific grammatical principle: subject-verb agreement. If the subject is complex, as in *the keys to the cabinet are/*is here*, this task requires implicitly learning some dependency structure, which some models are able to do. Evaluating a sentence representation system through acceptability classification makes it possible to expand the scope of experiments like these.

4 Related Work

There have been several prior attempts to use machine learning to learn a function that maps a sentence to a scalar acceptability score. These works vary considerably in the kinds of data used, the sources and natures of the acceptability judgments, and the role of labeled data.

Sources of Sentences It is possible to construct an acceptability corpus automatically if a method is found that can produce (roughly) unacceptable sentences. One approach is to programmatically generate *fake* sentences that are unlikely to be acceptable. Wagner et al. (2009) distort real sentences by, for example, deleting words, inserting words, or altering verbal inflection. Lau et al. (2016) use round-trip machine-translation from English into various languages and back. A second family of approaches take sentences from essays written by non-native speakers (Heilman et al., 2014). A third takes advantage of linguistics examples: Lawrence et al. (2000) and Lau et al. (2016) build datasets of 133 and 552 examples from syntax textbooks. CoLA scales up this line of work.

Sources of Judgments Wagner et al. (2009) label a sentence unacceptable if it has gone through one of their automatic distortion procedures. We take a similar approach in our auxiliary (real/fake) task (section 6). Heilman et al. (2014) and Lau et al. (2016) represent acceptability judgments on a continuous scale from 1 to 4, and average judgments across multiple speakers. Our labeling approach for CoLA follows that of Lawrence et al. (2000) in using already-annotated examples from the linguistics literature.

Grammatical Bias Prior work is inconsistent in the degree to which it gives models access to explicit information about English grammar. At one extreme, Lawrence et al. (2000) convert all their data to part-of-speech tags by hand, giving their model explicit access to these categories, which are not available to human learners. At the other extreme, Lau et al. (2016) use fully unsupervised methods, predicting acceptability as a function of probabilities assigned by unsupervised language models (LMs). We take care not to introduce linguistic bias in the form of grammatical annotations, though our models are partially supervised by example judgments.

5 CoLA

This paper introduces the Corpus of Linguistic Acceptability (CoLA), [6] a set of example sentences from the linguistics literature labeled for acceptability. Upon publication, CoLA will be made available

[6] CoLA can be downloaded here: `https://nyu-mll.github.io/CoLA/`

online, alongside source code for our baseline models, an interactive demo showing judgments by those models, and a leaderboard showing model performance on the test sets (using privately-held labels).

Sources We compile CoLA with the aim of representing a wide variety of phenomena of interest in theoretical linguistics. We draw examples from linguistics publications spanning a wide time period, a broad set of topics, and a range of target audiences. Table 2 enumerates our sources. By way of illustration, consider the three largest sources in the corpus: Kim & Sells (2008) is a recent undergraduate syntax textbook, Levin (1993) is a comprehensive reference detailing the lexical properties of thousands of verbs, and Ross (1967) is an influential dissertation on extraction and movement in English syntax.

Preparing the Data The corpus includes all usable examples from each source. We manually remove unacceptable examples falling into any of the excluded categories described in Section 2.3. The labels in the corpus are the original authors' acceptability judgments whenever possible. When examples appear with non-binary judgments (less than 3%), we either exclude them (for labels '?' or '#'), or label them unacceptable ('??' and '*?'). We also expand examples given with optional or alternate phrases into multiple data points. For example *Betsy buttered (*at) the toast* becomes *Betsy buttered the toast* and **Betsy buttered at the toast*.

In some cases, we change the content of examples slightly. To avoid irrelevant complications from out-of-vocabulary words, we restrict CoLA to the 100k most frequent words in the British National Corpus, and edit sentences as needed to remove words outside that set. For example, *That new handle unscrews easily* is replaced with *That new handle detaches easily* to avoid the out-of-vocabulary word *unscrews*. We make these alterations manually to preserve the author's stated intent, in this case selecting another verb that undergoes the middle voice alternation.

Finally we add content to examples that are not complete sentences, replacing, for example **The Bill's book* with **The Bill's book has a red cover*.

	N	%	Description
Adger (2003)	948	71.9	Syntax Textbook
Baltin (1982)	96	66.7	Movement
Baltin and Collins (2001)	880	66.7	Handbook
Bresnan (1973)	259	69.1	Comparatives
Carnie (2013)	870	80.3	Syntax Textbook
Culicover and Jackendoff (1999)	233	59.2	Comparatives
Dayal (1998)	179	75.4	Modality
Gazdar (1981)	110	65.5	Coordination
Goldberg and Jackendoff (2004)	106	77.4	Resultative
Kadmon and Landman (1993)	93	81.7	Negative Polarity
Kim and Sells (2008)	1965	71.2	Syntax Textbook
Levin (1993)	1459	69.0	Verb alternations
Miller (2002)	426	84.5	Syntax Textbook
Rappaport Hovav and Levin (2008)	151	69.5	Dative alternation
Ross (1967)	1029	61.8	Islands
Sag et al. (1985)	153	68.6	Coordination
Sportiche et al. (2013)	651	70.4	Syntax Textbook
In-Domain	**9515**	**71.3**	
Chung et al. (1995)	148	66.9	Sluicing
Collins (2005)	66	68.2	Passive
Jackendoff (1971)	94	67.0	Gapping
Sag (1997)	112	57.1	Relative clauses
Sag et al. (2003)	460	70.9	Syntax Textbook
Williams (1980)	169	76.3	Predication
Out-of-Domain	**1049**	**69.2**	
Total	**10657**	**70.5**	

Table 2: The contents of CoLA by source. *N* is the total number of examples. % is the percent of examples labeled acceptable. Sources listed above *In-Domain* are included in the training, development, and test sets, while those above *Out-of-Domain* appear only in the development and test sets.

Splitting the Data In addition to the train/development/test split used to control overfitting in standard benchmark datasets, CoLA is further divided into an in-domain set and an out-of-domain set, as specified in Table 2. The in-domain set is split three ways into a training set (8551 examples), a development set (527), and a test set (530), all drawn from the same 17 sources. The out-of-domain set is split into a development set (516) and a test set (533), both drawn from the same 6 sources. With two development and test sets we can monitor two types of overfitting during training: overfitting to the specific sentences in the training set (in-domain), and overfitting to the specific sources and phenomena represented in the training set (out-of-domain).

Label	Sentence	Source
0	The more books I ask to whom he will give, the more he reads.	Culicover and Jackendoff (1999)
1	I said that my father, he was tight as a hoot-owl.	Ross (1967)
1	The jeweller inscribed the ring with the name.	Levin (1993)
0	many evidence was provided.	Kim and Sells (2008)
1	They can sing.	Kim and Sells (2008)
1	The men would have been all working.	Baltin (1982)
0	Who do you think that will question Seamus first?	Carnie (2013)
0	Usually, any lion is majestic.	Dayal (1998)
1	The gardener planted roses in the garden.	Miller (2002)
1	I wrote Blair a letter, but I tore it up before I sent it.	Rappaport Hovav and Levin (2008)

Table 3: CoLA random sample, drawn from the in-domain training set. 1: acceptable, 0: unacceptable.

Human Performance We measure human performance on a subset of CoLA to set a reasonable upper bound for machine performance on acceptability classification, and to estimate the reproducibility of the judgments in CoLA. We obtained acceptability judgments from five linguistics PhD students on 200 sentences from CoLA, divided evenly between the in-domain and out-of-domain development sets. Results are shown in Table 4. Average accuracy across annotators is 86.1%, and average Matthews correlation coefficient (MCC)[7] is 0.697.

Selecting the majority decision from our annotators gives us a rough upper bound on human performance. These judgments agreed with CoLA's ratings on 87% of sentences with a MCC of 0.713. In other words, 13% of the labels in CoLA contradict the observed majority judgment. We identify several reasons for disagreements between our annotators and CoLA. Some sentences show copying errors which change the acceptability of the sentence or omit the original judgment. Other disagreements can be ascribed to unreliable judgments on the part of authors or a lack of context. We also measured our individual annotators' agreement with the aggregate rating, yielding an average agreement of 93%, and an average MCC of 0.852.

[7]Matthews correlation coefficient (Matthews, 1975) is our primary classification performance metric. It measures correlation on unbalanced binary classification tasks in range from -1 to 1, with any uninformed random guessing achieving an expected score of 0.

6 Experiments

6.1 Models

We train several neural network models to do acceptability classification on CoLA. At 10k sentences, CoLA is likely too small to train a low-bias learner like a recurrent neural network without additional prior knowledge. In similar low-resource settings, transfer learning with sentence embeddings has proven to be effective (Kiros et al., 2015; Conneau et al., 2017). We use transfer learning in all our models and train large sequence models on auxiliary tasks. In most experiments a large sentence encoder is trained on a real/fake discrimination task, and a lightweight multilayer perceptron classifier is trained on top to do acceptability classification over CoLA. Inspired by ELMo (Peters et al., 2018), we also experiment with using hidden states from an LSTM language model (LM) as word embeddings.

Real/Fake Pretraining Task We train sentence encoders to distinguish real and *fake* English sentences. The real data is drawn from the 100 million-token British National Corpus (BNC), and the fake data is a similar quantity automatically generated by two different strategies: We generate strings, e.g. (2-a), using an LSTM LM[8] trained on the BNC, and we manipulate sentences of the BNC, e.g. (2-b), by randomly permuting a subset of the words, keeping the other words *in situ*.

(2) a. either excessive tenure does not threaten a value to death.
b. what happened in to the empire early the tra-

[8]Trained to a word-level perplexity of 56.1.

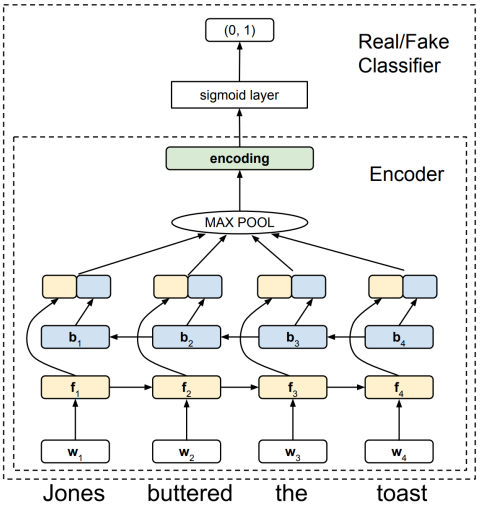

Figure 1: Real/Fake model (auxiliary pretraining setup). w_i = word embeddings, f_i = forward LSTM hidden state, b_i = backward LSTM hidden state.

ditional roman portrait?

This task is suitable because arbitrary numbers of labeled fake sentences can be generated without using any explicit knowledge of grammar in the process, and we expect that many of the same features are likely relevant to the real/fake task and the downstream acceptability task.

Real/Fake Encoder The real/fake model architecture is shown in Figure 1. A deep bidirectional LSTM reads a sequence of word embeddings. Then, following Conneau et al. (2017), the forward and backward hidden states for each time step are concatenated, and max-pooling over the sequence gives a sentence embedding. This is passed through a sigmoid output layer giving a scalar representing the probability that the sentence is real.

Acceptability Classifier Our acceptability classifier is a small two-layer network. The input is the fixed-length sentence embedding which is transferred from the real/fake encoder. This is passed through a $\tanh$ hidden layer followed by a sigmoid output layer. The encoder's weights are frozen during training on the acceptability task due the relatively small size of CoLA.

LM Encoder Further experiments use the LSTM LM described above as an encoder. The hidden states are transferred and an additional LSTM layer is trained on CoLA. As in Figure 1, the max pooling of the hidden states gives a sentence embedding which is passed to a sigmoid output layer.

Word Representations We experiment with several kinds of word representations: (i) We train word embeddings from scratch along with LSTMs on the language modeling or real/fake objectives. (ii) We use ELMo-style *contextualized* word embeddings from our trained LM. As in, ELMo (Peters et al., 2018), the representation for w_i here is a linear combination of the hidden states h_i^j for each layer j in an LSTM LM, though we depart from the original paper by using only a forward LM. (iii) In a more exploratory experiment, we also use pretrained 300-dimensional (6B) GloVe embeddings (Pennington et al., 2014). Note that these embeddings are trained on orders of magnitude more words than human learners ever see, limiting the possible interpretations of a positive result in this setting.

CBOW Baseline For a simple baseline, we train a continuous bag-of-words (CBOW) model directly on CoLA, using the sums of word embeddings from our best LSTM LM.

6.2 The Poverty of the Stimulus Revisited

Our experiments meet the design objectives laid out in Section 3.1 for testing the Poverty of the Stimulus Argument. The real/fake pretraining introduces no explicit grammatical bias into the sentence encoders; hence any linguistic features these models learn are acquired without instruction in what kinds of categories or structures are relevant to language understanding. The encoders are trained on 100-200 million tokens, which is within a factor of ten of the number of tokens human learners are exposed to during language acquisition (Hart and Risley, 1992).[9]

Unbiased sentence embeddings are the sole input to our acceptability classifiers. While the classifiers themselves are potentially biased from the roughly 9,000 expert-annotated sentences in the CoLA training set, practically all the linguistic knowledge in our models comes from the sequence model, which

[9] Hart and Risley (1992) find that children in affluent families are exposed to about 45 million tokens by age 4.

has several orders of magnitude more parameters and more training data. The acceptability classifier's role is primarily to extract from the sentence embedding the linguistic knowledge needed for the acceptability judgment. We control for any advantage CoLA training gives our models over naïve speakers by comparing their performance to that of trained linguists experienced with acceptability judgments. We also mitigate the impact of this training data by evaluating the model on the out-of-domain test set, in which it must reproduce judgments unrelated to those available to the model at training time.

6.3 Lau et al. Baselines

We compare our models with those of Lau et al. (2016). Their models obtain an acceptability prediction from unsupervised LMs by normalizing the LM output using one of several metrics. Following their recommendation, we use both the SLOR and Word LogProb Min-1 metrics.[10] Since these metrics produce unbounded scalar scores rather than probabilities or binary judgments, we fit a threshold to the outputs in order to use these models as acceptability classifiers. This is done with 10-fold cross-validation: we repeatedly find the optimum threshold for 90% of the model outputs and evaluate the remaining 10% with that threshold, until all the data have been evaluated. Following their methods, we train n-gram models on the BNC using their published code.[11] In place of their RNN LM, we use the same LSTM LM that we trained to generate sentences for the real/fake task.

6.4 Training details

All models are trained using PyTorch and optimized using Adam (Kingma and Ba, 2014). We train 20 LSTM LMs with from-scratch embeddings for up to 7 days or until completing 4 epochs without improving in development perplexity and select the best. Hyperparameters are chosen at random in these ranges: embedding size $\in$ [200, 600], hidden size $\in$ [600, 1200], number of layers $\in$ [1, 4], learning rate $\in$ [3×10^{-3}, 10^{-5}], dropout rate $\in \{0.2, 0.5\}$. We train 20 real/fake classifiers with from-scratch embeddings, 20 with GloVe, and 20 with ELMo-style embeddings for up to 7 days or until completing 4 epochs without improving in development MCC. Hyperparameters are chosen at random in these ranges: embedding size $\in$ [200, 600], hidden size $\in$ [600, 1400], number of layers $\in$ [1, 5], learning rate $\in$ [3×10^{-3}, 10^{-5}], dropout rate $\in \{0.2, 0.5\}$. We train 10 acceptability classifiers for each encoder until completing 20 epochs without improving in MCC on the CoLA development set. Hyperparameters are chosen at random in these ranges: hidden size $\in$ [20, 1200] and learning rate $\in$ [10^{-2}, 10^{-5}], dropout rate $\in \{0.2, 0.5\}$.

7 Results and Discussion

Table 4 shows our results. The best model is the real/fake model with ELMo-style embeddings. It achieves the highest MCC and accuracy both in-domain and out-of-domain by a large margin, outperforming even the models with access to GloVe.

All our models perform better than the unsupervised models of Lau et al. (2016) on both evaluation metrics on the in-domain test set. Out of domain, Lau et al.'s baselines offer the second-best results. Our models consistently perform worse out-of-domain than in-domain, with MCC dropping by as much as 50% in one case. Since Lau et al.'s baselines don't use the training set, they perform similarly in-domain and out-of-domain.

The sequence models consistently outperform the word order-independent CBOW baseline, indicating that the LSTM models are using word order for acceptability classification in a non-trivial way. In line with Lau et al.'s findings, the n-gram LM baselines are worse than the RNN LM. These results suggest that, unsurprisingly, LSTMs are better at capturing long-distance dependencies than n-gram models with a limited feature window.

Discussion Our LSTM models appear to be the best currently available low-bias learners for acceptability classification. Compared to humans, though, their absolute performance is underwhelming. We do not interpret this result as proof positive for the Poverty of the Stimulus Argument, though, as these

[10] Where s=sentence, $\mathrm{p_{LM}}(x)$ is the probability the LM assigns to string x, $\mathrm{p_u}(x)$ is the unigram probability of string x, and $|s|$ is the length of s: Word LP Min-1 = $\min\left\{-\frac{\log \mathrm{p_{LM}}(w)}{\log \mathrm{p_u}(w)},\ w \in s\right\}$ and SLOR = $\frac{\log \mathrm{p_{LM}}(s) - \log \mathrm{p_u}(s)}{|s|}$.

[11] https://github.com/jhlau/acceptability_prediction

Model	In-domain		Out-of-domain		Hyperparameters		
	Accuracy	MCC	Accuracy	MCC	Emb.	Enc.	H.
RNN LM Word LP Min-1	0.652	0.253	0.711	0.238	217	–	891
4-gram SLOR	0.642	0.223	0.645	0.042	–	–	–
3-gram SLOR	0.646	0.212	0.681	0.141	–	–	–
2-gram SLOR	0.590	0.162	0.707	0.180	–	–	–
CBOW Encoder w/ LM-Trained Embeddings	0.502	0.063	0.482	0.096	282	–	808
Real/Fake Encoder	0.723	0.261	0.679	0.186	505	1408	152
Real/Fake Encoder w/ GloVe Embeddings	0.706	0.300	0.608	0.135	300	1686	188
ELMo-Style Real/Fake Encoder	**0.772**	**0.341**	**0.732**	**0.281**	819	1056	1134
LM Encoder	0.726	0.278	0.651	0.155	217	819	629
Human Average	*0.850*	*0.644*	*0.872*	*0.738*	–	–	–
Human Aggregate	*0.870*	*0.695*	*0.910*	*0.815*	–	–	–

Table 4: Results for acceptability classification on the CoLA test set. *RNN LM* and n*-gram* are Lau et al.'s models. All models in the second section are acceptability classifiers trained on top the specified pretrained encoders. *Human Average* and *Human Aggregate* refer to the small human evaluations (Section 5). *Emb*: word embedding dim.. *Enc*: encoder state dim.. *H*: acceptability classifier hidden dim..

experiments reflect an early attempt at acceptability classification, and it is possible that more sophisticated low-bias models will decrease the performance gap substantially.

The supervised models see a substantial drop in performance from the in-domain test set to the out-of-domain test sets. This suggests that they've learned an acceptability model that is somewhat specialized to the phenomena in the training set, rather than the general English model one would expect. Addressing this problem will likely involve new forms of regularization to mitigate this overfitting and, more importantly, better pretraining strategies that can help the model learn the fundamental ingredients of grammaticality from *unlabeled* data.

We also measure the models' performance on individual sources in the in-domain development set. Performance is highly variable, with MCC ranging from 0.487 for the ELMo-style real/fake model on the Ross (1967) data, to 0.023 for the real/fake model with GloVe on Adger (2003).

8 Fine-Grained Analysis

Here, we run additional evaluations to probe whether our models are able to successfully learn grammatical generalizations. For these tests we generate five auxiliary datasets (described below) using simple rewrite grammars which target specific grammatical contrasts. The results from these experiments are shown in Table 5.

Unlike in CoLA, none of these judgments are meant to be difficult or controversial, and we expect that most humans could reach perfect accuracy. We also take care to make the test sentences as simple as possible to reduce classification errors unrelated to the target contrast. This is accomplished by limiting noun phrases to 1 or 2 words, and by using semantically related vocabulary items within examples.

8.1 Test Sets

Subject-Verb-Object This test set consists of 100 triples of subject, verb, and object each appearing in five permutations of (SVO, SOV, VSO, VOS, OVS).[12] The set of 100 triples is the Cartesian product product of three sets containing 10 subjects ({John, Nicole, ...}), 2 verbs ({read, wrote}), and 5 objects ({the book, the letter, ...}).

(3) John read the book. / *John the book read. / *read John the book. / *read the book John. / *the book read John.

[12]OSV is excluded because it does not have a clear acceptability rating. Examples such as "The book John read", can be interpreted as marginally acceptable sentences with topicalized subjects, or noun phrases with a relative clause modifier.

Model	SVO	Wh-Extraction	Causative	Subject-Verb	Reflexive
Real/Fake Encoder	0.381	0.184	0.463	0.098	0.043
Real/Fake Encoder w/ GloVe Embeddings	**0.988**	0.059	0.614	0.277	0.150
ELMo-Style Real/Fake Encoder	0.650	0.000	0.449	0.302	-0.020
LM Encoder	0.637	0.102	**0.633**	0.128	0.075
LSTM LM + Lau et al. Metrics	^{S}0.924	^{W}**0.601**	^{S}0.283	^{W}**0.599**	^{S}**0.521**

Table 5: Matthews Correlation Coefficient results for specific phenomena. ^{S}=SLOR, ^{W}=Word LP Min-1.

Wh-Extraction This test set consists of 260 pairs of contrasting examples, as in (4). This is to test (1) whether a model has learned that a wh-word must correspond to a gap somewhere in the sentence, and (2) whether the model can identify non-local dependencies up to three words away. The data contain 10 first names as subjects and 8 sets of verbs and related objects (5). Every compatible verb-object pair appears with every subject.

(4) a. What did John fry?
b. *What did John fry the potato?

(5) {{boil, fry}, {the egg, the potato}}

Causative-Inchoative Alternation This test set is based on a syntactic alternation conditioned by the lexical semantics of particular verbs. It contrasts verbs like *popped* which undergo the causative-inchoative alternation, with verbs like *blew* that do not. If *popped* is used transitively (6-a), the subject (*Kelly*) is an agent who causes the object (*the bubble*) to change states. Used intransitively (6-b), it is the subject (*the bubble*) that undergoes a change of state and the cause need not be specified (Levin, 1993). The test set includes 91 verb/object pairs, and each pair occurs in the two forms as in (6). 36 pairs allow the alternation, and the remaining 5 do not.

(6) a. Kelly popped/blew the bubble.
b. The bubble popped/*blew.

Subject-Verb Agreement This test set is generated from 13 subjects in singular and plural form crossed with 13 verbs in singular and plural form. This gives 169 quadruples as in (7).

(7) a. My friend has/*have to go.
b. My friends *has/have to go.

Reflexive-Antecedent Agreement This test set probes whether a model has learned that every reflexive pronouns must agree with an antecedent noun phrase in person, number, and gender. The dataset consists of a set of 4 verbs crossed with 6 subject pronouns and 6 reflexive pronouns, giving 144 sentences, only 1 out of 6 acceptable.

(8) I amused myself / *yourself / *herself / *himself / *ourselves / *themselves.

8.2 Results

The results in Table 5 show that LSTMs do make some systematic acceptability judgments as though they learn correct grammatical generalizations. Gross word order (*SVO* in Table 5) is especially easy for the models. The Real/Fake model with GloVe embeddings achieves near perfect correlation, suggesting that it systematically distinguishes gross word order. However, the remaining tests fall short of perfect performance.

Our models consistently outperform Lau et al.'s baselines on lexical semantics (*Causative*), judging more accurately whether a verb can undergo the causative-inchoative alternation. This may be due in part to the fact that our models are trained on CoLA which contains examples of similar alternations from Levin (1993).

Lau et al.'s baselines outperform our models on the remaining tests. It consistently identifies the long-distance dependency between a *wh*-word and its gap (*Wh-extraction*), while our models are near chance. Our models also perform relatively poorly on judgments involving agreement (*Singular/Pl*, *Reflexive*). While the poor absolute performance of these models is likely due to their inability to access sub-word morphological information, we have no working hypothesis to explain the relative success of the Lau et al. metrics.

9 Conclusion

This work offers resources and baselines for the study of semi-supervised machine learning for acceptability judgments. Most centrally, we introduce the first large-scale corpus of acceptability judgments, making it possible to train and evaluate modern neural networks on this task. In baseline experiments, we find that a network trained on our artificial real/fake task, combined with ELMo-style word representations, outperforms other available models, but remains far from human performance.

Much work remains to be done to implement the agenda described in Section 3. To provide stronger evidence on the Poverty of the Stimulus Argument, we hope for future work to test the performance of a broader range of effective candidate low-bias machine learning models, and to investigate how much can be gained by explicitly introducing specific forms of grammatical knowledge into models.

References

David Adger. 2003. *Core Syntax: A Minimalist Approach*. Oxford University Press Oxford.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Mark Baltin and Chris Collins, editors. 2001. *Handbook of Contemporary Syntactic Theory*. Blackwell Publishing Ltd.

Mark R Baltin. 1982. A landing site theory of movement rules. *Linguistic Inquiry*, 13(1):1–38.

Joan W Bresnan. 1973. Syntax of the comparative clause construction in English. *Linguistic Inquiry*, 4(3):275–343.

Andrew Carnie. 2013. *Syntax: A Generative Introduction*. John Wiley & Sons.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN-encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Noam Chomsky. 1957. *Syntactic Structures.* Mouton.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Sandra Chung, William A Ladusaw, and James McCloskey. 1995. Sluicing and logical form. *Natural Language Semantics*, 3(3):239–282.

Alexander Clark and Shalom Lappin. 2011. *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons.

Chris Collins. 2005. A smuggling approach to the passive in English. *Syntax*, 8(2):81–120.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector : Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Peter W Culicover and Ray Jackendoff. 1999. The view from the periphery: The English comparative correlative. *Linguistic Inquiry*, 30(4):543–571.

Veneeta Dayal. 1998. Any as inherently modal. *Linguistics and Philosophy*, 21(5):433–476.

Gerald Gazdar. 1981. Unbounded dependencies and coordinate structure. In *The Formal Complexity of Natural Language*, pages 183–226. Springer.

Edward Gibson and Evelina Fedorenko. 2010. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, 14(6):233–234.

Adele E Goldberg and Ray Jackendoff. 2004. The English resultative as a family of constructions. *Language*, 80(3):532–568.

Betty Hart and Todd R Risley. 1992. American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental Psychology*, 28(6):1096.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 174–180.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Ray S Jackendoff. 1971. Gapping and related rules. *Linguistic Inquiry*, 2(1):21–35.

Nirit Kadmon and Fred Landman. 1993. Any. *Linguistics and Philosophy*, 16(4):353–422, aug.

Jong-Bok Kim and Peter Sells. 2008. *English Syntax: An Introduction*. CSLI Publications.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2016. Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Steve Lawrence, C Lee Giles, and Sandiway Fong. 2000. Natural language grammatical inference with recurrent neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 12(1):126–140.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *arXiv preprint arXiv:1611.01368*.

Brian 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.

Jim Miller. 2002. *An Introduction to English Syntax*. Edinburgh University Press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Malka Rappaport Hovav and Beth Levin. 2008. The English dative alternation: The case for verb sensitivity. *Journal of Linguistics*, 44(1):129–167.

John Robert Ross. 1967. *Constraints on Variables in Syntax.* Ph.D. thesis, MIT.

Ivan A Sag, Gerald Gazdar, Thomas Wasow, and Steven Weisler. 1985. Coordination and how to distinguish categories. *Natural Language & Linguistic Theory*, 3(2):117–171.

Ivan A Sag, Thomas Wasow, and Emily M Bender. 2003. *Syntactic Theory: A Formal Introduction*. CSLI Publications, 2 edition.

Ivan A Sag. 1997. English relative clause constructions. *Journal of Linguistics*, 33(2):431–483.

Carson T Schütze. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534.

Dominique Sportiche, Hilda Koopman, and Edward Stabler. 2013. *An Introduction to Syntactic Analysis and Theory*. John Wiley & Sons.

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

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Joachim Wagner, Jennifer Foster, and Josef van Genabith. 2009. Judging grammaticality: Experiments in sentence classification. *CALICO Journal*, 26(3):474–490.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Edwin Williams. 1980. Predication. *Linguistic Inquiry*, 11(1):203–238.