Neural Network Acceptability Judgments

Alex Warstadt
New York University
warstadt@nyu.edu

Amanpreet Singh
New York University
Facebook AI Research*
amanpreet@nyu.edu

Samuel R. Bowman
New York University
bowman@nyu.edu

Abstract

This paper investigates the ability of artificial neural networks to judge the grammatical acceptability of a sentence, with the goal of testing their linguistic competence. 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 on acceptability classification, and find that our models outperform unsupervised models by Lau et al. (2016) on CoLA. Error-analysis on specific grammatical phenomena reveals that both Lau et al.'s models and ours learn systematic generalizations like subject-verb-object order. However, all models we test perform far below human level on a wide range of grammatical constructions.

1 Introduction

Artificial neural networks (ANNs) achieve a high degree of competence on many applied natural language understanding tasks, but it does not follow that they have knowledge of grammar. A key property of a human's linguistic competence is the ability to identify in one's native language, without formal training in grammar, a contrast in acceptability[1] between pairs of sentences like those in (1). *Acceptability judgments* like these are the primary behavioral measure that generative linguists use to observe humans' grammatical knowledge (Chomsky, 1957; Schütze, 1996).

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

We train neural networks to perform acceptability judgments—following work by Lawrence et al. (2000), Lau et al. (2016), and others—in order to evaluate their acquisition of the kinds of grammatical concepts linguists identify as central to human linguistic competence. This contributes to a growing effort to test ANNs' ability to make fine-grained grammatical distinctions (Linzen et al., 2016; Adi et al., 2017; Conneau et al., 2018; Ettinger et al., 2018; Marvin and Linzen, 2018). This research program seeks to provide new informative ways to evaluate ANN models popular with engineers. Furthermore, it has the potential to address foundational questions in theoretical linguistics by investigating how well unbiased learners can acquire grammatical knowledge.

In this paper we make four concrete contributions: (i) We introduce the Corpus of Linguistic Acceptability (CoLA), a collection of sentences from the linguistics literature with expert acceptability labels which, at over 10k examples, is by far the largest of its kind. (ii) 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. (iii) We analyze the impact of supervised training on acceptability classifiers by varying the domain and quantity of training data. (iv) We assess our models' performance on acceptability classification of specific linguistic phenomena. These experiments illustrate how acceptability classification and CoLA can give detailed insights into what grammatical knowledge typical neural network models can acquire. We find that our models do not show evidence of learning non-local dependencies related to agreement and questions, but do appear to acquire knowledge about basic subject-verb-object word order and verbal argument structure.

* Current affiliation. This work was completed when the author was at New York University.

[1] Following terminological conventions in linguistics, a sentence's *grammaticality* is determined by a grammatical formalism, and its *acceptability* is determined by introspective judgments of native speakers (Schütze, 1996).

Transactions of the Association for Computational Linguistics, vol. 7, pp. 625–641, 2019. https://doi.org/10.1162/tacl_a_00290
Action Editor: Alexander Clark. Submission batch: 2/2019; Revision batch: 6/2019; Published 9/2019.
© 2019 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

Resources CoLA can be downloaded from the corpus Web site.[2] The code for training our baselines 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 (unlabeled). Finally, CoLA is included in the GLUE benchmark[6] (Wang et al., 2018), which also hosts CoLA training data, unlabeled test data, and a leaderboard.

2 Acceptability Judgments

2.1 In Linguistics

Our investigation of acceptability classification builds on decades of established scientific knowledge in generative linguistics, where acceptability judgments are studied extensively. In his foundational work on generative syntax, Chomsky (1957) defines an empirically adequate grammar of a language L as one that generates all and only those strings of L which native speakers of L judge to be acceptable. Evaluating grammatical theories against native speaker judgments has been the dominant paradigm for research in generative syntax over the last sixty years (Schütze, 1996). Linguists generally provide evidence in the text of their papers in the form of constructed example sentences annotated with Boolean acceptability judgments from themselves or native speakers.

2.2 The Acceptability Classification Task

Although acceptability classification has been explored previously in computational linguistics, there is no standard approach to this task. Following common practice in generative linguistics, our study focuses on the Boolean acceptability classification task. This approach is also taken in earlier computational work on this task (Lawrence et al., 2000; Wagner et al., 2009; Linzen et al., 2016). By contrast, other computational work aims to model gradient acceptability judgments (Heilman et al., 2014; Lau et al., 2016). Though Lau et al. argue that acceptability judgments are gradient in nature, we consider Boolean judgments in published examples sufficient for our purposes, since linguists generally design these examples to be unambiguously acceptable or unacceptable.

Data sets for acceptability classification require a source of unacceptable sentences, which are not generally found in naturalistic speech or writing by native speakers. The sentences in CoLA consist entirely of examples from the linguistics literature. Lawrence et al. (2000) and Lau et al. (2016) build data sets similar in this respect. However, at over 10k sentences, CoLA is by far the largest data set of this kind, and represents the widest range of sources. Prior work in this area also obtains unacceptable sentences by programmatically generating 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. We also generate fake sentences to pre-train our baselines before further training on CoLA.

We see several advantages in using linguistics example sentences. First, they are labeled for acceptability by the authors, thereby simplifying the annotation process. Second, because linguists present examples to motivate arguments, these sentences isolate a particular grammatical construction while minimizing superfluous content. Hence, unacceptable sentences in CoLA tend to be maximally similar to acceptable sentences and are unacceptable for a single identifiable reason.

We note that Gibson and Fedorenko (2010) express concern about standard practices around acceptability judgments. They 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, 2017; Sprouse et al., 2013; Mahowald et al., 2016). We take no position on this general question, but perform a small human evaluation to gauge the reproducibility of the judgments in CoLA (Section 3).

2.3 The Role of Minimal Pairs

Acceptability judgments can alternatively be framed as a forced choice between *minimal pairs*, that is, pairs of minimally different sentences contrasting in acceptability as in (1), where the classifier or subject selects the sentence with greater (predicted) acceptability. This kind of

[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.

[6] https://gluebenchmark.com/tasks.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

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 informal classification of unnacceptable sentences, shown with their presence or absence in CoLA.

judgment has been taken as a standard for replicability of reported judgments in syntax articles (Sprouse and Almeida, 2012; Sprouse et al., 2013; Linzen and Oseki, 2018). It is also increasingly used in computational linguistics (Linzen et al., 2016; Marvin and Linzen, 2018; Futrell et al., 2018; Wilcox et al., 2018, 2019). This task is often used to evaluate language models because the outputted probabilities for a pair of minimally different sentences are directly comparable, while the output for a single sentence cannot be taken as a measure of acceptability without some kind of normalization (Lau et al., 2016).

We leave a comparison of this methodology with our own for future work. We settle on the single-sentence judgment task because it is directly comparable with methodology in generative linguistics. Although some work in theoretical linguists presents acceptability judgments as a ranking of two or more sentences (Schütze, 1996, pp. 77–81), Boolean judgments are still the norm, and the dominant current theories still make Boolean *predictions* about whether a sentence is or is not grammatical (Chomsky, 1995, pp. 12–16). Accordingly, CoLA, but not data sets based solely on preferences between minimal pairs, may be used to evaluate models' ability to make judgments that align with both native speaker judgments and the predictions of generative theories.

2.4 Defining (Un)acceptability

Not all linguistics examples are suitable for acceptability classification. Although 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) are interpretable, but in odd scenarios distinguishable from plausible scenarios only with access to real-world knowledge unrelated to grammar.

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 that 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. Because 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.

3 CoLA

This paper introduces the Corpus of Linguistic Acceptability (CoLA), a set of example sentences from the linguistics literature labeled for acceptability. See Table 3 for sample data. CoLA is available online, alongside source code for our baseline models, and a leaderboard showing model performance on test data using privately held labels (see footnotes 2–6 for links).

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

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 focusing on wh-movement and extraction islands 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 2.4. The labels in the corpus are the original authors' acceptability judgments whenever possible. When examples appear with non-Boolean judgments (this occurs in less than 3% of cases), we either exclude them (for labels '?' or '#'), or label them unacceptable ('??' and '*?'). We also expand examples 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 define acceptability classification as a sentence classification task. To ensure that all examples in CoLA are sentences, we augment fragmentary examples, replacing, for example, **The Bill's book* with **The Bill's book has a red cover.*

Splitting the data In addition to the train/development/test split used to control overfitting in standard benchmark data sets, CoLA is further divided into an in-domain set and an out-of-domain set, as specified in Table 2. The out-of-domain set is constructed to be about 10% the size of CoLA and to include sources of varying sizes, degrees of domain specificity, and time period.[7] The in-domain set is split three ways into training (8551 examples), development (527), and test sets (530), all drawn from the same 17 sources. The out-of-domain set is split into development (516) and a test sets (533), drawn from another 6 sources. We split CoLA in this way in order to monitor two

Source	N	%	Topic
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 number of sentences in a source. % is the percent of sentences labeled acceptable. Sources listed above *In-Domain* are included in the training, development, and test sets, whereas those above *Out-of-Domain* appear only in the development and test sets.

[7] In Section 6 we consider several alternate splits of CoLA.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

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

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).

Phenomena in CoLA CoLA has wide coverage of syntactic and semantic phenomena. To quantify the distribution of phenomena represented, we annotate the entire CoLA development set for the presence of constructions falling into 15 broad classes, of which 8 are discussed here, for brevity.[8] Briefly, *simple* labels sentences with no marked syntactic structures; *adjunct* labels sentences that contain adjuncts of nouns and verb phrases; *comp clause* labels sentences with embedded or complement clauses; *to-VP* labels sentences with non-finite embedded verb phrase; *arg altern* labels sentences with non-canonical argument structures such as passives; *binding* labels sentences with pronouns and binding phenomena; *question* labels sentences with interrogative clauses and relative clauses; and *violations* labels sentences with morphological or semantic violations, or an extra/missing word. The average sentence is labeled with 3.22 features.

Figure 1 shows the frequency of these 8 features in the development set. Argument alternations are the best represented phenomenon and appear in over 40% of sentences in this sample. This is due both to the high frequency of these constructions as well as the inclusion of several sources directly addressing this topic (Levin, 1993; Collins, 2005; Rappaport Hovav and Levin, 2008). Most

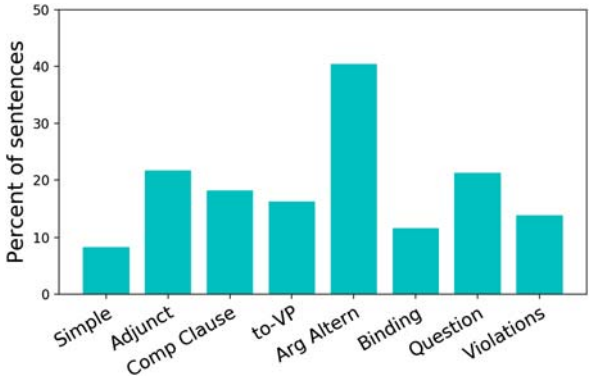


Figure 1: Frequencies of phenomenon types in the CoLA development set.

other constructions appear in about 10–20% of sentences, indicating that CoLA is fairly balanced according to this annotation scheme. There are likely biases in CoLA that other annotation schemes could detect. However, it is open to debate what a balanced data set for acceptability judgments should look like. There is no agreed-upon set of key phenomena in linguistics and any attempt to create one is likely to be controversial and overly simplistic. Furthermore, if such a set of phenomena did exist, the builders of a balanced data set must decide whether it should be balanced equally across phenomena, or weighted by either the frequency in broad coverage corpora of English or the number of distinguishing syntactic contrasts associated with each phenomenon. We assume that CoLA skews towards the latter, as a major goal of linguistics articles is to document key unique facts about some phenomenon without excessive repetition.

Human Performance We measure human performance on a subset of CoLA to set an approximate

[8]The annotated data also includes 63 fine-grained features. The annotated data is available for download on the CoLA website, and Warstadt and Bowman (2019) document annotation guidelines and conduct additional analysis.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

Model	Embeddings	Encoder Training	Classifier Training	In-domain Acc.	In-domain MCC	Out-of-domain Acc.	Out-of-domain MCC
CBOW	BNC	–	CoLA	0.502	0.063	0.482	0.096
LSTM LM WLPM	BNC	–	CoLA Thresh.	0.652	0.253	0.711	0.238
4-gram LM WLPM	–	–	CoLA Thresh.	0.474	0.000	0.645	0.042
3-gram LM WLPM	–	–	CoLA Thresh.	0.428	0.142	0.681	0.141
2-gram LM WLPM	–	–	CoLA Thresh.	0.452	0.094	0.707	0.180
Pooling Classifier	BNC	Real/Fake	Real/Fake	0.728	0.196	0.707	0.180
Pooling Classifier	GloVe	Real/Fake	Real/Fake	0.766	0.302	0.660	0.063
Pooling Classifier	ELMo-Style	Real/Fake	Real/Fake	0.758	0.265	0.702	0.177
Pooling Classifier	ELMo-Style	CoLA	CoLA	0.726	0.278	0.651	0.155
Pooling Classifier	BNC	Real/Fake	CoLA	0.723	0.261	0.679	0.186
Pooling Classifier	GloVe	Real/Fake	CoLA	0.706	0.300	0.608	0.135
Pooling Classifier	ELMo-Style	Real/Fake	CoLA	**0.772**	**0.341**	**0.732**	**0.281**
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. The first group is the CBOW baseline. The second group is the *LSTM* and n-*gram* LMs with Lau et al.'s metrics. The third group is pooling classifiers trained end-to-end on the real/fake objective. The fourth group is pooling classifiers with training on CoLA, mostly with encoders transferred from real/fake classifiers. The fifth group is the small human evaluations (Section 3). *CoLA-Thresh.* is threshold tuning on CoLA, and *WLPM* is Lau et al.'s Word LogProb Min-1 metric.

upper bound for machine performance on acceptability classification and to estimate the reproducibility of the judgments in CoLA. We have five linguistics PhD students, all native English speakers, perform a forced-choice single-sentence acceptability judgment task on 200 sentences from CoLA, divided evenly between the in-domain and out-of-domain development sets. These human judgments are available alongside on the corpus site.

Results appear in Table 4. Average annotator agreement with CoLA is 86.1%, and average Matthews Correlation Coefficient (MCC)[9] 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 an 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. Errors in character recognition in the source PDFs may produce artifacts which alter the acceptability of the sentence or omit the original judgment. Based on these 200 sampled sentences, we estimate such errors occur in 1–2% of CoLA sentences. Ascribing 2 percentage points of disagreement to such errors, the remaining 11 points can be ascribed to a lack of context or genuine variation between the dialect spoken by the original author and that spoken by the annotator.[10] We also measure our *individual* annotators' agreement with the aggregate rating, yielding an average pairwise agreement of 93%, and an average MCC of 0.852.

4 Experiments

We train several semi-supervised 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

[9]MCC (Matthews, 1975) is an evaluation metric for unbalanced binary classifiers. It is a special case of Pearson's r for Boolean variables, that is, it measures correlation of two Boolean distributions, giving a value between -1 and 1. On average, any two unrelated distributions will have an MCC of 0, regardless of class imbalance. By contrast, accuracy and F1 favor classifiers with a majority-class bias.

[10]We observe greater disagreement between human annotators and published judgments than Sprouse et al. (2013) do. As a reviewer points out, this may be due to the fact that Sprouse et al. measure agreement with minimal pairs of sentences using a forced choice task, which is more constrained and arguably easier than single sentence judgments.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

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). Our best model uses a transfer learning approach in which a large sentence encoder is trained on an unsupervised real/fake discrimination objective, and a lightweight multilayer perceptron classifier is trained on top to do acceptability classification over CoLA. It also uses contexualized word embeddings inspired by ELMo (Peters et al., 2018).

We compare our models to a continuous bag of words (CBOW) baseline, the unsupervised models proposed by Lau et al. (2016), and human performance. To make these comparisons more meaningful, we avoid giving our models distinct advantages over human learners by limiting the training data in two ways: (i) Aside from acceptability labels, our training has no grammatical annotation. (ii) Our large sentence encoders are limited to 100–200 million tokens of training data, which is within a factor of ten of the number of tokens human learners are exposed to during language acquisition (Hart and Risley, 1992).[11] We avoid training models on significantly more data because such models have a distinct advantage over the human learners we aim to match.

4.1 Preliminaries

Language model We use an LSTM language model (LSTM LM) at various stages in our experiments: (i) Several of our models use word embeddings or hidden states from the LM as input. (ii) The LM generates fake data for the real/fake task. (iii) The LM is an integral part of our implementation of the method proposed by Lau et al. (2016). We train the LM on the 100 million-token British National Corpus (BNC). It learns word embeddings from scratch for the 100k most frequent words in the BNC (with out of vocabulary words replaced by <unk>). We lowercase and tokenize the BNC data using NLTK (Bird and Loper, 2004). The LM achieves a word-level perplexity of 56.1 on the BNC.

Word representations We experiment with three styles of word representations: (i) We train a set of conventional fixed word embeddings as part of the training of the LM described above, which we refer to as *BNC embeddings*. (ii) We train *ELMo-style* contextualized word embeddings, which, following ELMo (Peters et al., 2018), represent w_i as 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) We also use the pretrained 300-dimensional (6B) *GloVe embeddings* from Pennington et al. (2014).[12]

Real/fake auxiliary task We train sentence encoders on a *real/fake task* in which the objective is to distinguish real sentences from the BNC and ''fake'' English sentences automatically generated by two strategies: (i) We sample strings (2-a) from the LSTM LM. (ii) We manipulate sentences of the BNC (2-b) by randomly permuting a subset of the words, keeping the other words *in situ*. Training data includes the entire BNC and an equal amount of fake data. We lowercase and tokenize all real/fake data and replace out of vocabulary words as in LM training.

(2) a. either excessive tenure does not threaten a value to death.
b. what happened in to the empire early the traditional roman portrait?

We choose this task 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 relevant to both the real/fake task and the downstream acceptability task.

4.2 Baselines

Pooling classifier Our real/fake classifiers and acceptability classifiers use an architecture we refer to as a *pooling classifier*, which is based on Conneau et al. (2017). As illustrated in Figure 2, the pooling classifier consists of two parts: (i) a sentence encoder which reduces variable-length sequences of tokens into fixed-length *sentence embeddings*, and (ii) a lightweight classifier which outputs a classification based on the sentence embedding. In the sentence encoder, a deep bidirectional LSTM reads a sequence of word embeddings;

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

[12] Results with models that use these GloVe embeddings are less immediately comparable with human performance results, since GloVe is trained on several orders of magnitude more text than humans see during language acquisition.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

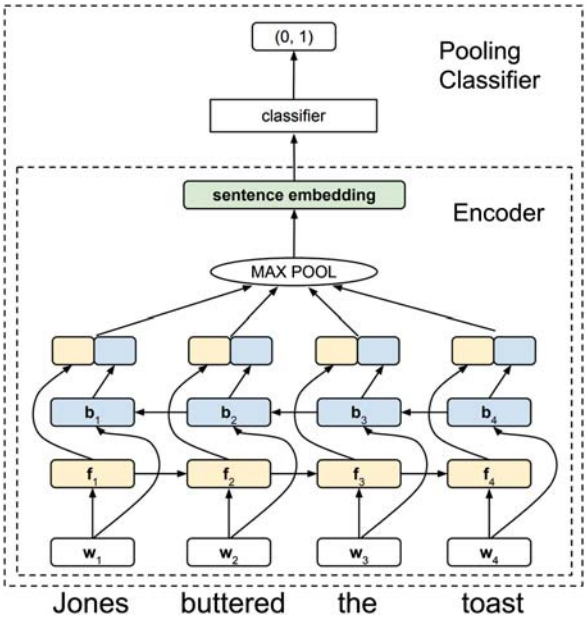


Figure 2: Architecture for the pooling classifier models. w_i = word embeddings, f_i = forward LSTM hidden state, b_i = backward LSTM hidden state.

then the forward and backward hidden states for each time step are concatenated, and max-pooling over the sequence gives a sentence embedding. In the classifier, the sentence embedding is passed through a sigmoid output layer (optionally preceded by a single hidden layer) giving a scalar representing the probability of a positive classification (either the sentence is real or acceptable, depending on the task).

We train several variations of pooling classifiers, as shown in Table 4. First, we train classifiers end-to-end on the real/fake task, varying the style of word embedding. The classifier portion consists only of a single softmax layer. We evaluate these classifiers on CoLA without CoLA training.

Second, we train pooling classifiers entirely on CoLA. We test only ELMo-style embeddings here because, unlike BNC and GloVe embeddings, they include robust contextual information about the entire sequence, eliminating the need for training a large LSTM on CoLA alone.

Third, we transfer features learned from the real/fake task to classifiers trained on CoLA. Specifically, we freeze the weights of the sentence encoder portion of the real/fake classifiers, and train new classifiers on CoLA using the sentence embeddings as input. For these experiments, in addition to a sigmoid layer, the classifier has an additional hidden $tanh$ layer to compensates for the fact that the sentence encoder is not fine-tuned on CoLA.

Lau et al. (2016) We compare our models to 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 the *Word LogProb Min-1* metric.[13] Because this metric produces unbounded scalar scores rather than probabilities or Boolean 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 on the CoLA test set: 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.[14] In place of their RNN LM, we use the same LSTM LM that we use to generate sentences for the real/fake task.

CBOW For a simple baseline, we train a CBOW model directly on CoLA. We pass the sum of BNC word embeddings for the sentence to a multilayer perceptron with a single hidden layer.

4.3 Training Details

All neural network models are implemented in 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 four epochs without improving in development perplexity and select the best checkpoint. Hyperparameters for each experiment 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 select the model with best performance for use in further experiments.

We train 20 pooling classifiers end-to-end on real/fake data with BNC embeddings, 20 with

[13] Where s=sentence, $\mathrm{p_{LM}}(x)$ is the probability the LM assigns to string x and $\mathrm{p_u}(x)$ is the unigram probability of string x: Word LP Min-1 = $\min\left\{-\frac{\log \mathrm{p_{LM}}(w)}{\log \mathrm{p_u}(w)},\ w \in s\right\}$. Lau et al. also obtain strong results with the *SLOR* metric. We also calculate results with *SLOR* but find them to be slightly worse overall, though not universally. We do not report these results, but they are available upon request.

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

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

GloVe, and 20 with ELMo-style embeddings for up to 7 days or until completing four epochs without improving in development MCC. We train 20 pooling classifiers end-to-end on CoLA using ELMo-style embeddings. Hyperparameters are chosen at random in these ranges: embedding size $\in$ [200, 600], hidden size $\in$ [500, 1500], number of layers $\in [1, 5]$, learning rate $\in [3\times10^{-3}, 10^{-5}]$, dropout rate $\in \{0.2, 0.5\}$.

For transfer learning experiments, we extract and freeze the weights from the encoders from the 5 best real/fake classifiers with BNC, GloVe, and ELMo-style embeddings, each. For every encoder, we train 10 classifiers on CoLA until completing 20 epochs without improving in MCC on the 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 $\in \{0.2, 0.5\}$.

For our single best model—a pooling classifier with ELMo-style embeddings, an encoder with real/fake training, and a classifier with CoLA training—the embedding size (i.e., LM hidden size) is 819 dimensions, the real/fake encoder hidden layer size is 528 dimensions, and the acceptability classifier hidden layer size is 1134.

5 Results and Discussion

Table 4 shows the results of the best run from each experiment. The best model overall 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 models with real/fake encoders and CoLA training 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. Because Lau et al.'s baselines don't use the training set, they perform similarly in-domain and out-of-domain. Real/fake classifiers without any additional training on CoLA tend to perform significantly worse than their counterparts with CoLA supervision.

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 LSTM LM. This result is expected given that n-gram models, but not LSTMs, have a limited feature window.

Discussion Of the models we have tested, LSTMs are the most effective low-bias learners for acceptability classification. Compared with humans, though, their absolute performance is underwhelming. This indicates to us that whereas the ANNs we study can acquire substantial knowledge of grammar, their linguistic competence is far from rivaling that of humans.

Our models with unsupervised pretraining have an advantage over similar models without pretraining. This finding aligns with the conclusions of Peters et al. (2018). We see this effect with both the LM pretraining for our ELMo-style embeddings real/fake pretraining for our sentence encoders. Unsurprisingly, the unsupervised Lau et al. models and real/fake classifiers are not as effective as models trained on CoLA. However, they far outperform random guessing and the CBOW baseline, indicating that even purely unsupervised models acquire significant knowledge of grammar.

The supervised models universally see a substantial drop in performance from the in-domain test set to the out-of-domain test set. This suggests that they have specialized somewhat to the phenomena in the training set, rather than English grammar in a fully general way as one would hope for. Addressing this problem will likely involve new forms of regularization to mitigate this overfitting and, more importantly, new pretraining strategies that can help the model better learn the fundamental ingredients of grammaticality from unlabeled data.

6 CoLA Design Experiments

The results in the previous section highlight the effects of pretraining, but give little insight into how the labeled training data in CoLA impacts classifier performance. To quantify the impact of CoLA training, we conduct two additional experiments: First, we measure how the amount of training data impacts model performance on the CoLA development set. Second, we investigate how the specific contents of the in-domain and out-of-domain sets impact model generalization.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

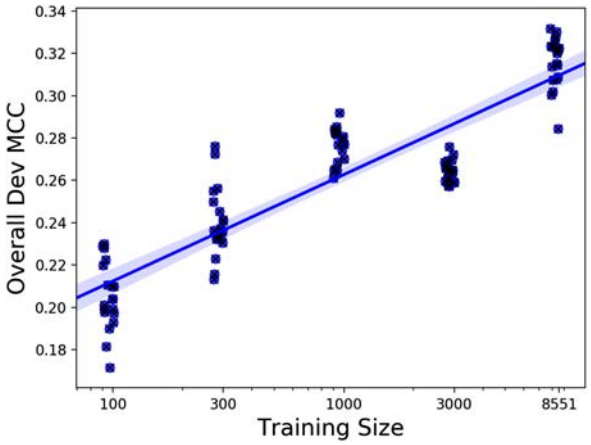


Figure 3: Results on the CoLA development set as a function of the number of training examples, with line of best fit and 95% confidence interval. Random x jitter added.

Training set size In this experiment, we vary the amount of training data seen by our acceptability classifiers. We construct alternate training sets of sizes 100, 300, 1000, and 3000 by randomly downsampling the 8551-example CoLA training set. Then, for each training set we train classifiers with 20 restarts using the best performing ELMo-style real/fake encoder, and evaluate on the entire development set. Figure 3 plots the results. As training data increases from 100 to 8551 sentences, we see approximately log-linear improvements in performance. The small decrease in performance between 1000 and 3000 sentences is likely an artifact of the random downsampling.

From these results we draw two main conclusions: First, it appears that increasing the amount of training data in CoLA by an order of magnitude may significantly benefit our models. Second, much of what our models learn from CoLA can be learned from as few as 300 training examples. This suggests that CoLA training is not teaching our models specific facts about acceptability as much as teaching them to use existing grammatical knowledge from the sentence encoders.

Splitting CoLA Our results in Table 4 show that our models' performance drops noticeably when tested on out-of-domain sentences from publications not represented in the training data. In this experiment, we investigate different splits of CoLA into in-domain and out-of-domain to test the degree to which the decrease in performance on out-of-domain sentences is a stable property of these models, or simply an artifact of the particular publications represented in the out-of-domain set (as described in Section 3).

The splits are constructed by randomly selecting sources from the 23 sources from CoLA to hold out until the sum of their sizes exceeds 750. This gives out-of-domain set sizes ranging from 789 to 1539, consisting of 2 to 6 sources. CoLA's original out-of-domain set contains 1049 examples and 6 sources. Development and test sets are constructed by randomly splitting the out-of-domain data in half, and randomly selecting an approximately equal number of in-domain sentences. For each training set we train classifiers with 20 restarts using the encoder from the best performing ELMo-style real/fake classifier.

In Table 5, we report the average test performance over 20 restarts. We conclude that the domain difference between two samples of sources in CoLA is generally a meaningful one for these models. This is especially so for the original split, where average in-domain MCC is 0.125 greater than out-of-domain MCC, close to the maximum observed difference of 0.162. By contrast, in one case average out-of-domain performance was actually better. This tells us that the particular nature of the sources in each domain has a large effect on what our models learn.

7 Phenomenon-Specific Analysis

In addition to testing the general grammatical knowledge of low-bias learners, acceptability classification can be used to probe models' knowledge of particular linguistic phenomena. We analyze our baselines' performance by phenomenon using two methods: First, we break down their performance on CoLA based on the different constructions present in the target sentences. Second, we evaluate them on controlled test sets targeting specific grammatical contrasts.

7.1 CoLA Performance by Phenomenon

In this error analysis, we study performance on CoLA as a function of the syntactic features of the individual sentences, using the 8 features described in Section 3. We train classifiers with 20 restarts using the best performing ELMo-style real/fake encoder. For each feature, we measure the MCC of our models on only those sentences with that feature.

Figure 4 shows the mean MCC over 20 restarts for each feature. Unsurprisingly, syntactically

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

Split	In-Domain		Out-of-Domain		Overall		Out Sources	Out N
	Acc.	MCC	Acc.	MCC	Acc.	MCC		
orig.	0.701	0.348	0.620	0.223	0.660	0.285	C05, J71, S97, CLC95, W80, SWB04	1049
1	0.729	0.357	0.632	0.195	0.680	0.275	BC01, B73	1139
2	0.700	0.319	0.666	0.188	0.683	0.255	KL93, SGWW85, W80, D98, B73, G81	853
3	0.708	0.333	0.659	0.284	0.684	0.307	AD03, D98, G81	1237
4	0.663	0.243	0.673	0.267	0.668	0.252	B82, SWB04, CJ99	789
5	0.720	0.349	0.671	0.285	0.696	0.315	M02, BC01, CJ99	1539

Table 5: Results for 5 different splits of CoLA and the original split into in-domain and out-of-domain. All results are averages over 20 restarts. Out N is the number of out-of-domain sentences. Sources are abbreviated by authors' last initial and year; full citations for each source are shown in Table 2.

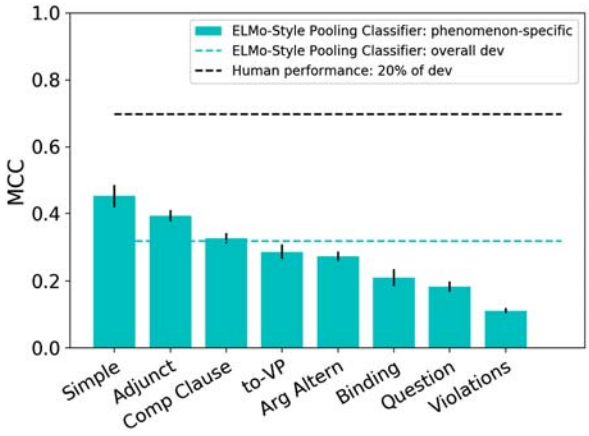


Figure 4: Performance on phenomenon-specific subsets of the CoLA development set. Results are the mean over 20 random restarts, with error bars ± 1 STD. Lines show mean performance on the entire dev set and mean human performance on 200 dev set sentences. *Simple*: no marked syntactic structures. *Adjunct*: adjuncts of nouns and verb phrases. *Comp Clause*: embedded or complement clauses. *to-VP*: non-finite embedded verb phrase. *Binding*: pronouns and binding phenomena. *Question*: questions and relative clauses. *Violations*: morphological or semantic violations, or an extra/missing word.

simple sentences are easier than average, but unexpectedly sentences with adjuncts are as well. Sentences with complement clauses, embedded VPs, and argument alternations are about as hard as the average sentence in CoLA. Although these constructions can be complex, they also occur with very high frequency. Sentences with binding and violations, including morphological violations, are among the hardest. We also find that our models perform poorly on sentences with question-like syntax. This difficulty is likely due to long-distance dependencies in these sentences.

7.2 Targeted Test Sets

Here, we run additional evaluations to probe whether our models can reliably classify sets of sentences that target a single grammatical contrast. This kind of evaluation can give insight into what kinds of grammatical features our models do and do not acquire easily. Using data generation techniques inspired by Ettinger et al. (2016), we build five auxiliary data sets (described below) using simple rewrite grammars which target specific grammatical contrasts.

Unlike in CoLA, none of these judgments are meant to be difficult or subtle, 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. Specifically, we limit noun phrases to 1 or 2 words and use semantically related vocabulary items within examples.

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).[15] The set of 100 triples is the Cartesian product product of three sets containing 10 subjects ({John, Bo, ...}), 2 verbs ({read, wrote}), and 5 objects ({the book, the letter, ...}).

(3) a. Bo read the book. b. *Bo the book read.
c. *read Bo the book. d. *read the book Bo.
e. *the book read Bo.

Wh-Extraction This test set consists of 260 pairs of contrasting examples, as in (4). This is

[15]OSV is excluded because it does not yield a clear acceptability rating. Examples such as ''The book John read'', can be interpreted as marginally acceptable sentences with topicalized subjects, or as acceptable noun phrases (rather than sentences) with relative clause modifiers.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

Model	Emb.	Enc.	Class.	SVO	Wh	Causative	SV Agr.	Reflexive
LSTM LM WLPM	BNC	–	CoLA Th.	0.801	**0.601**	0.270	**0.599**	**0.152**
Pooling	ELMo-St.	CoLA	CoLA	0.637	0.102	**0.633**	0.128	0.075
Pooling	BNC	R/F	CoLA	0.381	0.184	0.463	0.098	0.043
Pooling	GloVe	R/F	CoLA	**0.988**	0.059	0.614	0.277	0.150
Pooling	ELMo-St.	R/F	CoLA	0.650	0.000	0.449	0.302	-0.020

Table 6: MCC results for specific phenomena. *Emb*. is model embedding style; *Enc*. is model encoder training, *Class*. is model classifier training. *R/F* is real/fake, *ELMo-St.* is ELMo-style, and *CoLA-Th.* is threshold tuning on CoLA. *LSTM LM WLPM* is the LM with Lau et al. metrics Word LP Min-1.

to test (i) whether a model has learned that a wh-word must correspond to a gap in the sentence, and (ii) 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 direct 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). Thirty-six pairs allow the alternation, and the remaining 55 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 Example (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 data set 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.

Results The results from these experiments are given in Table 6. Our models' performance on these test sets is mixed. They make some systematic acceptability judgments that reflect correct grammatical generalizations. Some models are very effective at judging violations in gross word order (*SVO* in Table 6). The pooling classifier with GloVe embeddings achieves near perfect correlation, suggesting that it systematically uses gross word order. However, the remaining tests yield much poorer 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 receive supervision from CoLA, in which argument alternations are well represented (see Figure 1).

Lau et al.'s baseline outperforms our models in some cases. The LSTM LM with the Word LP Min-1 metric is the only model that can reliably identify the non-local dependency between a *wh*-word and its gap (*Wh*). It also performs relatively better on judgments involving agreement (*SV Agr.*). All models struggle on the *Reflexive* examples.

The poor performance of our models on contrasts involving agreement ((*SV Agr.*) and *Reflexive*) is surprising in light of findings by Linzen et al. (2016) that LSTMs can identify agreement errors easily even without access to sub-word

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

information. We speculate that this is due to under-representation of the relevant examples in CoLA. We estimate that morphological violations make up about 6% of examples in CoLA (about half of the *Violations* in Figure 1).

8 Motivation & Related Work

We see two chief motivations that guide work on acceptability classification with ANNs by us and by others: First, more fine-grained evaluation tools may accelerate work on general-purpose neural network modules for sentence understanding. Second, studying the linguistic competence of ANNs bears on foundational questions in linguistics about the learnability of grammar.

Fine-grained evaluation of ANNs The question of how well ANNs learn fine-grained grammatical distinctions has been the subject of much recent work. One method is to train models to perform probing tasks which target a construction of interest. Examples of such tasks are to determine whether the sentence is in active or passive voice (Shi et al., 2016), whether the subject is singular or plural (Conneau et al., 2018), or whether a given token is under the scope of negation (Ettinger et al., 2018). In each case, the authors use these tasks to compare the performance of reusable sentence embeddings.

Acceptability classification can be used to target many of the same grammatical constructions as probing tasks. For instance, an acceptability classifier that can reliably distinguish between pairs of sentences as in (9) must have implicit knowledge of the whether the subject of a sentence is singular or plural (in the first case) and whether the token *ever* is under the scope of negation. These exact experiments have been conducted by Linzen et al. (2016) and Marvin and Linzen (2018), respectively, although these works differ from our approach in that they do not evaluate domain general acceptability classifiers on these contrasts.

(9) a. The key is/*are on the table.
    b. Betsy hasn't/*has ever been to France.

Acceptability classification also enables certain kinds of investigations not possible with probing tasks. A single acceptability classifier can be trained to identify numerous unrelated contrasts. This is generally not possible with probing tasks, because the classes are tied to specific grammatical concepts. Acceptability classification also encourages direct comparison between ANN and human linguistic competence because, unlike many probing tasks, it can be easily performed by native speakers without linguistic training. Finally acceptability classifiers and generative grammars share a common objective, namely to predict the well-formedness of all and only those strings of the language that are acceptable to native speakers. Accordingly, it is straightforward to draw parallels between acceptability classifiers and established work in generative linguistics.

The poverty of the stimulus Research on acceptability classification can also be brought to bear on a foundational question in linguistic theory: The extent to which human linguistic competence is learned or innate. The influential *argument from the poverty of the stimulus* (APS) holds that the extent of human linguistic competence cannot be explained by purely domain general learning mechanisms and that humans must be born with a Universal Grammar which imparts specific knowledge of grammatical universals to the child and makes learning possible (Chomsky, 1965). While the APS has been subject to much criticism (Pullum and Scholz, 2002), it remains a foundation of much of contemporary linguistics.

In the setting of machine learning, the APS predicts that any artificial leaner trained with no prior knowledge of the principles of syntax and no more data than a human child sees must fail to make acceptability judgments with human-level accuracy (Clark and Lappin, 2011). If linguistically uninformed neural network models achieve human-level performance on specific phenomena or on a domain-general data set like CoLA, this would be clear evidence limiting the scope of phenomena for which the APS can hold.

However, acceptability classification alone cannot evaluate aspects of ANNs' linguistic competence against humans' in every relevant way. For example, Berwick et al. (2011) note that native speakers can easily recognize that, in *Bo is easy to please*, Bo is the entity being *pleased*, while in *Bo is eager to please*, Bo is the one who does the *pleasing*. Because the acceptability judgments in CoLA are reading-independent (see Table 1), they cannot be used to probe whether ANNs understand these distinctions.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

We wish to stress that the success of supervised acceptability classifiers like the ones we train cannot falsify the APS, because unacceptable examples play no apparent role in child language acquisition. While unsupervised acceptability classification could do so, more work is needed to find methods for extracting reliable Boolean acceptability judgments from unsupervised language models. Our approach of fitting a threshold to the models of Lau et al. (2016) gives encouraging results, but these models are ultimately not as effective as supervised models. An alternative adopted by Linzen et al. (2016) and Marvin and Linzen (2018) is to evaluate whether language models' assign higher probability to the acceptable sentence in a minimal pair. However, this forced choice minimal pair task, as discussed in Section 2.3, cannot be applied to CoLA, which does not exclusively contain minimal pairs.

Still, we maintain that our approach is a valuable step in the direction of evaluating the APS. Our results strongly suggest that grammatically unbiased sentence embeddings and contextualized word embeddings have non-trivial implicit knowledge of grammar before supervised training on CoLA. As our experiments in Section 6 show, a significant portion of what these models learn from CoLA can be learned from relatively little acceptability judgment data (as few as 300 sentences, of which fewer than 100 are unacceptable). Furthermore, the real/fake encoders and ELMo-style embeddings are trained on a quantity of data comparable to what human learners are exposed to. Given the rapid pace of development of new robust sentence embeddings, we expect to see increasingly human-like acceptability judgments from powerful neural networks in coming years, though with an eye towards evaluating the APS, future work should continue to investigate acceptability classifiers with unsupervised methods and restricted training resources.

9 Conclusion

This work offers resources and baselines for the study of semi-supervised machine learning for acceptability judgments. Most centrally, we introduce CoLA, 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 8. There is much untapped potential in the acceptability classification task as a fine-grained evaluation tool and as a test of the Poverty of the Stimulus Argument. We hope for future work to test the performance of a broader range of new effective low-bias machine learning models on CoLA, and to investigate further what grammatical principles these models do and do not learn.

Acknowledgments

This project has benefited from help and feedback at various stages from Chris Barker, Pablo Gonzalez, Shalom Lappin, Omer Levy, Marie-Catherine de Marneffe, Alex Wang, Alexander Clark, everyone in the Deep Learning in Semantics seminar at NYU, and three anonymous TACL reviewers. This project has benefited from financial support to S.B. by Google, Tencent Holdings, and Samsung Research. This material is based upon work supported by the National Science Foundation under grant no. 1850208. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

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

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR Conference Track*. Tulon.

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

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

Robert C. Berwick, Paul Pietroski, Beracah Yankama, and Noam Chomsky. 2011. Poverty of the stimulus revisited. *Cognitive Science*, 35(7):1207–1242.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, page 31.

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.

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

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

Noam Chomsky. 1995. *The Minimalist Program*. 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, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

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. In *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2126–2136.

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.

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.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

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.

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.

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

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

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.

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. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Tal Linzen and Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: A Journal of General Linguistics*, 3(1), 100.

Kyle Mahowald, Peter Graff, Jeremy Hartman, and Edward Gibson. 2016. SNAP judgments: A small n acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language*, 92(3):619–635.

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.

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*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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.

Geoffrey K. Pullum and Barbara C. Scholz. 2002. Empirical assessment of stimulus poverty arguments. *Linguistic Review*, 18(1–2):9–50.

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

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*, 2nd ed. CSLI Publications.

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.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024

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 and Diogo Almeida. 2017. Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable. *Behavioral and Brain Sciences*, 40:e311.

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.

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

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.

Alex Warstadt and Samuel R. Bowman. 2019. Grammatical analysis of pretrained sentence encoders with acceptability judgments. *arXiv preprint arXiv:1901.03438*.

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.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019. Structural supervision improves learning of non-local grammatical dependencies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3302–3312.

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

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00290 by guest on 28 March 2024