Targeted Syntactic Evaluation of Language Models

Rebecca Marvin
Department of Computer Science
Johns Hopkins University
becky@jhu.edu

Tal Linzen
Department of Cognitive Science
Johns Hopkins University
tal.linzen@jhu.edu

Abstract

We present a dataset for evaluating the grammaticality of the predictions of a language model. We automatically construct a large number of minimally different pairs of English sentences, each consisting of a grammatical and an ungrammatical sentence. The sentence pairs represent different variations of structure-sensitive phenomena: subject-verb agreement, reflexive anaphora and negative polarity items. We expect a language model to assign a higher probability to the grammatical sentence than the ungrammatical one. In an experiment using this data set, an LSTM language model performed poorly on many of the constructions. Multi-task training with a syntactic objective (CCG supertagging) improved the LSTM's accuracy, but a large gap remained between its performance and the accuracy of human participants recruited online. This suggests that there is considerable room for improvement over LSTMs in capturing syntax in a language model.

1 Introduction

A language model (LM) defines a probability distribution over sequences of words. Recent technological advances have led to an explosion of neural network-based LM architectures. The most popular ones are based on recurrent neural networks (RNNs) (Elman, 1990; Mikolov et al., 2010), in particular Long Short-Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997). While a large number of alternative architectures have been proposed in the past few years, LSTMs are still highly competitive (Melis et al., 2018).

Language models are typically evaluated using perplexity: it is considered desirable for an LM to assign a high probability to held-out data from the same corpus as the training data. This measure conflates multiple sources of success (or failure) in predicting the next word: common collocations, semantics, pragmatics, syntax, and so on. The quality of the **syntactic** predictions made by the LM is arguably particularly difficult to measure using perplexity: since most sentences are grammatically simple and most words can be predicted from their local context, perplexity rewards LMs primarily for collocational and semantic predictions.

We propose to supplement perplexity with a metric that assesses whether the probability distribution defined by the model conforms to the grammar of the language. Following previous work (Lau et al., 2017; Linzen et al., 2016; Gulordava et al., 2018), we suggest that given two sentences that differ minimally from each other, one of which is grammatical and the other which is not, it is desirable for the model to assign a higher probability to the grammatical one.

The value of this approach can be illustrated with a recent study by Tran et al. (2018), where a standard LSTM language model was compared to an attention-only LM without recurrence (Vaswani et al., 2017). Although the attention-only model had somewhat better perplexity on the validation set, when the models were tested specifically on challenging subject-verb agreement dependencies, the attention-only model made three times as many errors as the LSTM. In other words, the LSTM learned more robust syntactic representations, but this advantage was not reflected in its average perplexity on the corpus, since syntactically challenging sentences are relatively infrequent.

Previous work on targeted syntactic evaluation of language models has identified syntactically challenging sentences in corpora (Linzen et al., 2016; Gulordava et al., 2018). While evaluation on naturally occurring examples is appealing, this approach has its limitations (see Section 2). In particular, syntactically challenging examples are sparsely represented in a corpus, their identifica-

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1192–1202
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

tion requires a clean parsed corpus, and naturally occurring sentences are difficult to control for confounds. We contrast the naturalistic approach with a constructed dataset, which allows us to examine a much larger range of specific grammatical phenomena than has been possible before. We use templates to automatically create our test sentences, making it possible to generate a large test set while maintaining experimental control over our materials as well as a balanced number of examples of each phenomenon.

We test three LMs on the data set we develop: an *n*-gram baseline, an RNN LM trained on an unannotated corpus, and an RNN LM trained on a multitask objective: language modeling and Combinatory Categorial Grammar (CCG) supertagging (Bangalore and Joshi, 1999). We also conduct a human experiment using the same materials. The *n*-gram baseline largely performed at chance, suggesting that good performance on the task requires syntactic representations. The RNN LMs performed well on simple cases, but struggled on more complex ones. Multi-task training with a supervised syntactic objective improved the performance of the RNN, but it was still much weaker than humans. This suggests that our data set is challenging, especially when explicit syntactic supervision is not available, and can therefore motivate richer language modeling architectures.

2 Overview of the approach

2.1 Grammaticality and LM probability

How should grammaticality be captured in the probability distribution defined by an LM? The most extreme position would be that a language model should assign a probability of zero to ungrammatical sentences. For most applications, some degree of error tolerance is desirable, and it is not practical to assign a sentence a probability of exactly zero.[1] Following Linzen et al. (2016) and Gulordava et al. (2018), our desideratum for the language model is more modest: if two closely matched sentence differ only in their grammaticality, the probability of the grammatical sentence should be higher than the probability of the ungrammatical one. For example, the following minimal pair illustrates the fact that third-person present English verbs agree with the number of their subject:

(1) *Simple agreement*:
  a. The author laughs.
  b. *The author laugh.

We expect the probability of (1a) to be higher than the probability of (1b). Previous work has simplified this setting further by comparing the probability that the LM assigns to a single word that is the locus of ungrammaticality. In (1), for example, the LM would be fed the first two words of the sentence, and would be considered successful on the task if it predicts $P(\textit{laughs}) > P(\textit{laugh})$.

The prediction setting is only applicable when the locus of ungrammaticality is a single word, rather than, say, the interaction between two words; moreover, the information needed to make the grammaticality decision needs to be available in the left context of the locus of grammaticality. These conditions do not always hold. Negative polarity items (NPIs), for example, are words like *any* and *ever* that can only be used in the scope of negation.[2] The grammaticality of placing a particular quantifier in the beginning of the sentences in (2) depends on whether the sentence contains an NPI later on:

(2) *Simple NPI:*
  a. No students have ever lived here.
  b. *Most students have ever lived here.

It would not be possible to compare these two sentences using the prediction task. In the current paper, we use the more general setting and compare the probability of the two complete sentences.

2.2 Data set construction

Previous work has used syntactically complex sentences identified from a parsed corpus. This approach has several limitations. If the corpus is automatically parsed, the risk of a parse error increases with the complexity of the construction (Bender et al., 2011). If the test set is restricted to sentences with gold parses, it can be difficult or impossible to find a sufficient number of examples of syntactically challenging cases. Moreover, using naturally occurring sentences can introduce

[1] Nor is it possible to have a threshold ϵ such that all grammatical sentences have probability higher than ϵ and all ungrammatical sentences have probability lower than ϵ, for the simple reason that there is an infinite number of grammatical sentences (Lau et al., 2017).

[2] In practice, the conditions that govern the distribution of NPIs are much more complicated, but this first approximation will suffice for the present purposes. For a review, see Giannakidou (2011).

confounds that may complicate the interpretation of the experiments (Ettinger et al., 2018).

To circumvent these issues, we use templates to automatically construct a large number of English sentence pairs (∼350,000). Our data set includes three phenomena that linguists consider to be sensitive to hierarchical syntactic structure (Everaert et al., 2015; Xiang et al., 2009): subject-verb agreement (described in detail in Sections 4.1 and 4.2), reflexive anaphora (Section 4.3) and negative polarity items (Section 4.4).

The templates can be described using non-recursive context-free grammars. We specify the preterminal symbols that make up a syntactic construction and have different terminal symbols that those preterminals could be mapped to. For example, the template for the simple agreement construction illustrated in (1) consists of the following rules:

(3) a. Simple agreement → D MS MV
b. D → *the*
c. MS → {*author*, *pilot*, . . .}
d. MV → {*laughs*, *smiles*, . . .}

We generate all possible combinations of the terminals. The Supplementary Materials provide a full description of all our templates.[3]

While these examples are somewhat artificial, our goal is to isolate the syntactic capabilities of the model; it is in fact beneficial to minimize the semantic or collocational cues that can be used to identify the grammatical sentence. Gulordava et al. took this approach further and constructed "colorless green ideas" test cases by substituting random content words into sentences from a corpus. We take a more moderate position and avoid combinations that are very implausible or violate selectional restrictions (e.g., *the apple laughs*). We do this by having separate templates for animate and inanimate subjects and verbs so that the resulting sentences are always reasonably plausible.

3 Related work

Targeted evaluation: LM evaluation data sets using challenging prediction tasks have been proposed in the context of semantics and discourse comprehension (Zweig and Burges, 2011; Paperno et al., 2016). Evaluation sets consisting of challenging syntactic constructions have been constructed for parser evaluation (Rimell et al., 2009; Nivre et al., 2010; Bender et al., 2011), and minimal pair approaches have been proposed for evaluating image captioning (Shekhar et al., 2017) and machine translation systems (Sennrich, 2017), but no data sets exist that target a range of syntactic constructions for language model evaluation.

Acceptability judgments: Lau et al. (2017) compared the ability of different LMs to predict graded human acceptability judgments. The forced-choice approach used in the current paper has been shown to be effective in human acceptability judgment experiments (Sprouse and Almeida, 2017). In some early work, neural networks were trained explicitly to predict acceptability judgments (Lawrence et al., 1996; Allen and Seidenberg, 1999); Post (2011) likewise trained a classifier on top of a parser to predict grammaticality. Warstadt et al. (2018) use a transfer learning approach, where an unsupervised model is fine-tuned on acceptability prediction. Our work differs from those studies in that we do not advocate providing any explicit grammaticality signal to the LM at any point ("no negative evidence").

Syntax in LMs: There have been several proposals over the years to incorporate explicit syntax into LMs to overcome the inability of *n*-gram LMs to model long-distance dependencies (Jurafsky et al., 1995; Roark, 2001; Pauls and Klein, 2012). While RNN language models can in principle model longer dependencies (Mikolov et al., 2010; Linzen et al., 2016), in practice it can still be beneficial to inject syntax into the model. This can be done by combining it with a supervised parser (Dyer et al., 2016) or other multi-task learning objectives (Enguehard et al., 2017). Our work is orthogonal to this area of research, but can be seen as providing a potential opportunity to underscore the advantage of such syntax-infused models.

4 Data set composition

This section describes all of the types of sentence pairs included in our data set, which include examples of subject-verb agreement (Sections 4.1 and 4.2), reflexive anaphoras (Section 4.3) and negative polarity items (Section 4.4).

[3] The code, the data set and the Supplementary Materials can be found at `https://github.com/BeckyMarvin/LM_syneval`.

4.1 Subject-verb agreement

Determining the correct number of the verb is trivial in examples such as (1) above, in which the sentence only contains a single noun. By contrast, in cases where there are multiple nouns in the sentence, identifying which of them is the subject of a given verb requires understanding the structure of the sentence. In particular, the relevant subject is not necessarily the first noun of the sentence:

(4) *Agreement in a sentential complement:*
a. The bankers knew the officer smiles.
b. *The bankers knew the officer smile.

Here the verb *smiles* needs to agree with the embedded subject *officer* rather than the main clause subject *bankers*. The subject is also not necessarily the most recent noun before the verb: when the subject is modified by a phrase, a distracting noun ("attractor") often intervenes in the linear order of the sentence between the head of the subject and the verb. Two examples of such modifiers are prepositional phrases and relative clauses (RCs):

(5) *Agreement across a prepositional phrase:*
a. The farmer near the parents smiles.
b. *The farmer near the parents smile.

(6) *Agreement across a subject relative clause:*
a. The officers that love the skater smile.
b. *The officers that love the skater smiles.

We include all four possible configurations of noun number for each type of minimal pair; for (5), these would be:[4]

(7) a. The farmer near the parent smiles/*smile.
b. The farmer near the parents smiles/*smile.
c. The farmers near the parent smile/*smiles.
d. The farmers near the parents smile/*smiles.

Sentences where the two nouns conflict in number are expected to be more challenging, but interpretable errors may certainly occur even when they do not. For example, the model may use the heuristic that sentences with multiple nouns are likely to have a plural verb (a heuristic that would be effective for coordination); alternatively, it might prefer singular verbs to plural ones regardless of whether the subject is singular or plural, simply because the singular form of the verb is more frequent.

Next, in verb phrase (VP) coordination, both of the verbs need to agree with the subject:

(8) *Short VP coordination:*
a. The senator smiles and laughs.
b. *The senator smiles and laugh.

We had both singular and plural subjects. The number of the verb immediately adjacent to the subject was always grammatical. This problem can in principle be solved with a trigram model (*smiles and laughs* is likely to be a more frequent trigram than *smiles and laugh*); to address this potential concern, we also included a coordination condition with a longer dependency:

(9) *Long VP coordination:*
The manager writes in a journal every day and likes/*like to watch television shows.

4.2 Agreement and object relative clauses

We go into greater depth in object relative clauses, which most clearly require a hierarchical representation. In (10) and (11), the model needs to be able to distinguish the embedded subject (*parents*) from the main clause subject (*farmer*) when making its predictions:

(10) *Agreement across an object relative clause:*
a. The farmer that the parents love swims.
b. *The farmer that the parents love swim.

(11) *Agreement in an object relative clause:*
a. The farmer that the parents love swims.
b. *The farmer that the parents loves swims.

In keeping with the minimal pair approach, we never introduce two agreement errors at the same time: either the embedded verb or the main verb is incorrectly inflected, but not both.

We include a number of variations on the pattern in (11). First, we delete the relativizer *that*, with the hypothesis that the absence of an overt cue to structure will make the task more difficult:

(12) The farmer the parents love/*loves swims.

In another condition, we replace the main subject with an inanimate noun and keep the embed-

[4] The slash notation indicates the word that differs between the grammatical and ungrammatical sentence; for example, in (7a), the full sentence pair would be:

(i) a. The farmer near the parent smiles.
b. *The farmer near the parent smile.

ded subject animate. We base this manipulation on human experimental work showing that similar nouns (for example, two animate nouns) are more likely to cause confusion during comprehension than dissimilar nouns, such as an animate and an inanimate noun (Van Dyke, 2007):

(13) The movies that the author likes are/*is good.

For a complete list of all the types of minimal pairs we include, see the Supplementary Materials.

4.3 Reflexive anaphora

A reflexive pronoun such as *himself* needs to have an antecedent from which it derives its interpretation. The pronoun needs to agree in number (and gender) with its antecedent:

(14) *Simple reflexive*:
  a. The senators embarrassed themselves.
  b. *The senators embarrassed herself.

There are structural conditions on the nouns to which a reflexive pronoun can be bound. One of these conditions requires the antecedent to be in the same clause as the reflexive pronoun. For example, (15b) cannot refer to a context in which *the pilot* embarrassed *the bankers*:

(15) *Reflexive in a sentential complement*:
  a. The bankers thought the pilot embarrassed himself.
  b. *The bankers thought the pilot embarrassed themselves.

Likewise, in the following minimal pair, sentence (16b) is ungrammatical, because the reflexive pronoun *themselves*, which is part of the main clause, cannot be bound to the noun phrase *the architects*, which is inside an embedded clause:

(16) *Reflexive across an object relative clause:*
  a. The manager that the architects like doubted himself.
  b. *The manager that the architects like doubted themselves.

4.4 Negative polarity items

Negative polarity items, introduced in example (2) above, are words that (to a first approximation) need to occur in the context of negation. Crucially for the purposes of the present work, the scope of negation is structurally defined. In particular the negative noun phrase needs to c-command the NPI: the syntactic non-terminal node that dominates the negative noun phrase must also dominate the NPI. This is the case in (17a), but not in (17b), where the negative noun phrase is too deep in the tree to c-command the NPI *ever* (Xiang et al., 2009; Everaert et al., 2015).

(17) *NPI across a relative clause:*
  a. No authors that the security guards like have ever been famous.
  b. *The authors that no security guards like have ever been famous.

All of the nouns and verbs in the NPI cases were plural. As in some of the agreement cases, we included a variant of (17) in which the subject was inanimate.

5 Experimental setup

To show how our challenge set can be used to evaluate the syntactic performance of LMs, we trained three LMs with increasing levels of syntactic sophistication. All of the LMs were trained on a 90 million word subset of Wikipedia (Gulordava et al., 2018). Our *n*-gram LM and LSTM LM do not require annotated data. The third model is also an LSTM LM, but it requires syntactically annotated data (CCG supertags).

***N*-gram model:** We trained a 5-gram model on the same 90M word corpus using the SRILM toolkit (Stolcke, 2002) which backs off to smaller *n*-grams using Kneser-Ney smoothing.

Single-task RNN: The RNN LM had two layers of 650 LSTM units, a batch size of 128, a dropout rate of 0.2, and a learning rate of 20.0, and was trained for 40 epochs (following the hyperparameters of Gulordava et al. 2018).

Multi-task RNN: In multi-task learning, the system is trained to optimize an objective function that combines the objective functions of several tasks. We combine language modeling with CCG supertagging, a task that predicts for each word in the sentence its CCG supertag (Bangalore and Joshi, 1999; Lewis et al., 2016). We simply sum the two objective functions with equal weights (Enguehard et al., 2017). Early stopping in this model is based on the combined loss on language modeling and supertagging. Supertags provide a large amount of syntactic information about the word; the sequence of supertags of a sentence strongly constrains the possible parses of the sentence. We use supertagging as a "scaffold" task (Swayamdipta et al., 2017): our goal is not to produce a competitive supertagger, but to induce better syntactic representations, which would then lead to improved language modeling. We used CCG-Bank (Hockenmaier and Steedman, 2007) as our CCG corpus.

Human evaluation: We designed a human experiment on Amazon Mechanical Turk that mirrored the task that was given to the LMs: both versions of a minimal pair were shown on the screen at the same time, and participants were asked to judge which one of them was more acceptable (for details, see the Supplementary Materials). We emphasize that we do not see human performance on complex syntactic dependencies as setting an upper bound on the performance that we should expect from an LM. There is a rich literature showing that humans make mistakes such as subject-verb agreement errors; in fact, most of the phenomena we test were inspired by work in psycholinguistics that studies these errors (Bock and Miller, 1991; Phillips et al., 2011). At the same time, while we do not see a reason not to aspire for 100% accuracy, we are interested in comparing LM and human errors: if the errors are similar, the two systems may be using similar representations.

6 Results

Local agreement: The overall accuracy per condition can be seen in Table 1. The *n*-gram LM's accuracy was only 79% for simple agreement and agreement in a sentential complement, both of which can be solved entirely using local context. This is because not all subject and verb combinations in our materials appeared verbatim in the 90M word training corpus; for those combinations, the model fell back on unigram probabilities, which in this context amounts to selecting the more frequent form of the verb.

Both RNNs performed much better than the *n*-gram model on the simple agreement case (single-task: 94%; multi-task: 100%), reflecting these models' ability to generalize beyond the specific bigrams that occurred in the corpus. Accuracy on agreement in a sentential complement was also very high (single-task: 99%; multi-task: 93%). This indicates that the RNNs do not rely on the heuristic whereby the first noun of the sentence is likely to be its subject. They did slightly worse but still very well on short VP coordination (both 90%); this dependency is also local, albeit across the word *and*.

Non-local agreement: The accuracy of the *n*-gram model on non-local dependencies (long VP coordination and agreement across a phrase or a clause) was very close to 50%. This suggests that local collocational information is not useful in these conditions. The single-task RNN also performed much more poorly on these conditions than on the local agreement conditions, though for the most part its accuracy was better than chance. Humans did worse on these dependencies as well, but their accuracy did not drop as sharply as the RNNs' (human accuracies ranged from 82% to 88%). In most of these cases, multi-task learning was very helpful; for example, accuracy in long VP coordination increased from 61% to 81%. Still, both RNNs performed poorly on agreement across an object RC, especially without *that*, whereas humans performed comparably on all non-local dependencies.

Agreement inside an object RC: This case is particularly interesting, because this dependency is purely local (see (11)), and the interference is from the distant sentence-initial noun. Although this configuration is similar to the sentential complement case, performance was worse both in RNNs and humans. However, RNNs performed *better* than humans, at least when the sentence included the overt relativizer *that*. This suggests that interference is sensitive to proximity in RNNs but to syntactic status in humans — humans appear to be confusing the main clause subject and the embedded subject (Wagers et al., 2009).

Reflexive anaphora: The RNNs' performance was significantly worse on simple reflexives (83%) than on simple agreement (94%), and did not differ between the single-task and multi-task models. By contrast, human performance did not differ between subject-verb agreement and reflexive anaphoras. The surprisingly poor performance for this adjacent dependency seems to be due to an asymmetry in accuracy between *himself* and *themselves* on the one hand (100% accuracy in the multi-task RNN) and *herself* on the other hand (49% accuracy).[5] Accuracy was very low for all

[5] This may be because *himself* and *themselves* are significantly more frequent than *herself*, and consequently the num-

	RNN	Multitask	*n*-gram	Humans	# sents
SUBJECT-VERB AGREEMENT:					
Simple	0.94	1.00	0.79	0.96	280
In a sentential complement	0.99	0.93	0.79	0.93	3360
Short VP coordination	0.90	0.90	0.51	0.94	1680
Long VP coordination	0.61	0.81	0.50	0.82	800
Across a prepositional phrase	0.57	0.69	0.50	0.85	44800
Across a subject relative clause	0.56	0.74	0.50	0.88	22400
Across an object relative clause	0.50	0.57	0.50	0.85	44800
Across an object relative (no *that*)	0.52	0.52	0.50	0.82	44800
In an object relative clause	0.84	0.89	0.50	0.78	44800
In an object relative (no *that*)	0.71	0.81	0.50	0.79	44800
REFLEXIVE ANAPHORA:					
Simple	0.83	0.86	0.50	0.96	560
In a sentential complement	0.86	0.83	0.50	0.91	6720
Across a relative clause	0.55	0.56	0.50	0.87	44800
NEGATIVE POLARITY ITEMS:					
Simple	0.40	0.48	0.06	0.98	792
Across a relative clause	0.41	0.73	0.60	0.81	31680

Table 1: Overall accuracies for the LSTMs, *n*-gram model and humans on each test case.

pronouns in the structurally complex case in which the dependency was across a relative clause (55% compared to 87% in humans).

NPIs: The dependency in simple NPIs spans only four words, so the *n*-gram model could in principle capture it. In practice, the *n*-gram model systematically selected the wrong answer, suggesting that it backed off to comparing the bigrams *no students* and *most students*, the first of which is presumably less frequent. Surprisingly, the *n*-gram model's accuracy was higher than 50% on NPIs across a relative clause, a dependency that spans more than five words. In this case, the bigrams *that the* and *the chef* (for example) happen to be more frequent than the *that no* and *no chef*. This difference was apparently strong enough to make up for the low-frequency bigram at the start of the sentence.

The RNNs did poorly on this task. The accuracy of the single-task model was around 40%. The multi-task did somewhat better on the simple NPIs (48%) and much better on the NPIs across a relative clause (73%). At the same time, an examination of the plot of log probability of each word in a sentence (Figure A.1 in the Supplementary Materials) suggests that the single-task RNN is in fact able to differentiate between the grammatical and ungrammatical sentences when it reaches the NPI, but this difference does not offset the overall probability advantage of the ungrammatical sentence (which is likely due to non-grammatical collocational factors). In any case, the fact that the *n*-gram baseline did not perform at chance suggests that there are non-syntactic cues to this task, complicating the interpretation of the performance of other LMs.

Perplexity: The perplexity of the *n*-gram model on the Wikipedia test data was 157.5, much higher than the perplexity of the single-task RNN (78.65) and the multi-task RNN (61.10). In other words, perplexity tracked accuracy on our syntactic data set – an unsatisfying outcome given our goal of dissociating perplexity and our syntactic evaluation method, but an expected one given that each model was conditioned on richer information than the previous one. In previous work, perplexity and syntactic judgment accuracy have been found to be partly dissociable (Kuncoro et al., 2018; Tran et al., 2018).

Lexical variation and frequency: There was considerable lexical variation in the results; we have mentioned the surprising asymmetry between *himself* and *herself* above. As another case study, we examine variation in the results of the simple agreement condition in the single-

ber representation learned for *herself* was not robust. Another possibility is that gender bias reduces the probability of an anaphoric relation between *herself* and words such as *surgeon* (Rudinger et al., 2018).

Main subject	Embedded subject	Single-task	Multi-task	Humans	Example sentence
Across an objective relative clause:					
Singular	Singular	0.83	0.77	0.96	The author that the minister likes laughs/*laugh.
Singular	Plural	0.51	0.30	0.90	The author that the ministers like laughs/*laugh.
Plural	Singular	0.18	0.53	0.77	The authors that the minister likes laugh/*laughs.
Plural	Plural	0.50	0.73	0.80	The authors that the ministers like laugh/*laughs.
Within an objective relative clause:					
Singular	Singular	0.73	0.92	0.94	The author that the minister likes/*like laughs.
Singular	Plural	0.91	0.81	0.72	The author that the ministers like/*likes laugh.
Plural	Singular	0.81	0.97	0.73	The authors that the minister likes/*like laugh.
Plural	Plural	0.87	0.84	0.76	The authors that the ministers like/*likes laugh.

Table 2: Accuracy within and across an object relative clause (only in the cases in which the main subject was animate and the relativizer *that* was present). The subject that the verb is expected to agree with is underlined.

task RNN. Accuracy varied by verb, ranging from *is* and *are*, which had 100% accuracy, to *swims*, where accuracy was only 60% (recall that average accuracy was 94%). This may be a frequency effect: either the LM is learning less robust number representations for infrequent verbs, or the tail of the distribution over the vocabulary is more fragile during word prediction. Pauls and Klein (2012) propose normalizing for unigram frequency when deriving acceptability judgments from an LM. Our preliminary experiments with this method did not significantly improve overall performance; regardless of the effectiveness of this method, such corrections should arguably not be necessary in an LM that adequately captures grammaticality.

7 Case study: agreement and object relative clauses

The overall results in Table 1 were averaged over all of the possible number configurations within each condition. In this section, we take a closer look at agreement in sentences with an object RC (see Table 2). This kind of finer-grained analysis helps explain the cases in which the LMs are failing, and might reveal some of the patterns or heuristics the LMs are using.

Performance in agreement across an object RC was poor. Both RNNs made attraction errors: they often preferred the verb that agreed in number with the irrelevant embedded subject to the verb that agreed with the correct main subject. The multi-task RNN showed greater symmetry between the simpler singular/singular and plural/plural cases, whereas the single-task RNN performed poorly even in these cases, often preferring a singular verb when both subjects were plural. This default preference for singular verbs matches the behavior of younger children (Franck et al., 2004).

Performance in agreement within an object RC was better; still, the single-task RNN made the most errors when both subjects were singular, perhaps due to a heuristic in which a sentence with multiple subjects is likely to have a plural verb (as in coordination sentences). By contrast, the multi-task model seemed to have a general bias towards singular subjects in this condition. Incidentally, the human results with object RCs were also unexpected: while attraction errors when the two subjects differ in number are to be expected (Wagers et al., 2009), our participants made a sizable number of errors even when both subjects were plural.

Despite the generally poor performance in object RCs, Figures A.2 and A.3 in the Supplementary Materials show that the single-task RNN is typically assigning a higher probability to the grammatical word of a minimal pair than to the ungrammatical word.

8 Discussion

We have described a template-based data set for targeted syntactic evaluation of language models. The data set consists of pairs of sentences that are matched except for their grammaticality; we consider a language model to capture the relevant aspects of the grammar of the language if it assigns a higher probability to the grammatical sentence than to the ungrammatical one.

An RNN language model performed very well on local subject-verb agreement dependencies, significantly outperforming an *n*-gram baseline.

This suggests that the task is a viable evaluation strategy. Even on simple cases, however, the RNN's accuracy was sensitive to the particular lexical items that occurred in the sentence; this would not be expected if its syntactic representations were fully abstract. The RNN's performance degraded markedly on non-local dependencies, approaching chance levels on agreement across an object relative clause. Multi-task training with a syntactic objective (CCG supertagging) mitigated this drop in performance for some but not all of the dependencies we tested. We conjecture that the benefits of the inductive bias conferred by multi-task learning will be amplified when the amount of training data is limited.

Our results contrast with the results of Gulordava et al. (2018), who obtained a prediction accuracy of 81% on English sentences from their test corpus and 74% on constructed sentences modeled after sentences from the corpus. It is likely that our sentences are more syntactically challenging than the ones they were able to find in the relatively small manually annotated treebank they used.

One limitation of our approach is that it is not always clear what constitutes a minimal grammaticality contrast. In the subject-verb agreement case, the contrast was clear: the two present-tense forms of the verb, e.g., *laugh* vs. *laughs*. Our NPI manipulations, on the other hand, were less successful: the members of the contrasts differed not only in their syntactic structure but also in low-level *n*-gram probabilities, making the performance on this particular contrast harder to interpret.

We emphasize that the goal of this article was not to advocate for LSTMs in particular as an effective architecture for modeling syntax; indeed, our results show that LSTM language models are far from matching naive annotators' performance on this task, let alone performing at 100% accuracy. We hope that our data set, and future extensions to other phenomena and languages, will make it possible to measure progress in syntactic language modeling and will lead to better understanding of the syntactic generalizations captured by language models.

9 Acknowledgments

We would like to thank Ming Xiang for sharing materials from human experiments that inspired many of our test cases. We also thank Brian Roark and the JHU Computational Psycholinguistics lab for discussion, and Brian Leonard for help conducting the human experiment.

References

Joseph Allen and Mark S. Seidenberg. 1999. The emergence of grammaticality in connectionist networks. In Brian MacWhinney, editor, *Emergentist approaches to language: Proceedings of the 28th Carnegie symposium on cognition*, pages 115–151. Mahwah, NJ: Lawrence Erlbaum Associates.

Srinivas Bangalore and Aravind K. Joshi. 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265.

Emily M. Bender, Dan Flickinger, Stephan Oepen, and Yi Zhang. 2011. Parser evaluation over local and non-local deep dependencies in a large corpus. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 397–408. Association for Computational Linguistics.

Kathryn Bock and Carol A. Miller. 1991. Broken agreement. *Cognitive Psychology*, 23(1):45–93.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209. Association for Computational Linguistics.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Émile Enguehard, Yoav Goldberg, and Tal Linzen. 2017. Exploring the syntactic abilities of RNNs with multi-task learning. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 3–14. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Martin B. H. Everaert, Marinus A. C. Huybregts, Noam Chomsky, Robert C. Berwick, and Johan J. Bolhuis. 2015. Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, 19(12):729–743.

Julie Franck, Stephany Cronel-Ohayon, Laurence Chillier, Ulrich H. Frauenfelder, Cornelia Hamann, Luigi Rizzi, and Pascal Zesiger. 2004. Normal and pathological development of subject–verb agreement in speech production: A study on French children. *Journal of Neurolinguistics*, 17(2-3):147–180.

Anastasia Giannakidou. 2011. Negative and positive polarity items: Variation, licensing, and compositionality. In *Semantics: An international handbook of natural language meaning*, pages 1660–1712. Berlin: Mouton de Gruyter.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205. Association for Computational Linguistics.

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

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Daniel Jurafsky, Chuck Wooters, Jonathan Segal, Andreas Stolcke, Eric Fosler, G Tajchaman, and Nelson Morgan. 1995. Using a stochastic context-free grammar as a language model for speech recognition. In *Proceedings of ICASSP*, volume 1, pages 189–192. IEEE.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436. Association for Computational Linguistics.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, (5):1202–1247.

Steve Lawrence, Lee C. Giles, and Santliway Fong. 1996. Can recurrent neural networks learn natural language grammars? In *IEEE International Conference on Neural Networks*, volume 4, pages 1853–1858.

Mike Lewis, Kenton Lee, and Luke Zettlemoyer. 2016. LSTM CCG parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 221–231. Association for Computational Linguistics.

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.

Gábor Melis, Chris Dyer, and Phil Blunsom. 2018. On the state of the art of evaluation in neural language models. In *Proceedings of the International Conference on Learning Representations*.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, pages 1045–1048, Makuhari, Chiba, Japan.

Joakim Nivre, Laura Rimell, Ryan McDonald, and Carlos Gómez Rodríguez. 2010. Evaluation of dependency parsers on unbounded dependencies. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 833–841. Coling 2010 Organizing Committee.

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

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968. Association for Computational Linguistics.

Colin Phillips, Matthew W. Wagers, and Ellen F. Lau. 2011. Grammatical illusions and selective fallibility in real-time language comprehension. In Jeffrey T. Runner, editor, *Experiments at the Interfaces, Syntax and Semantics 37*, pages 153–186. Bingley, U.K.: Emerald.

Matt Post. 2011. Judging grammaticality with tree substitution grammar derivations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 217–222. Association for Computational Linguistics.

Laura Rimell, Stephen Clark, and Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 813–821. Association for Computational Linguistics.

Brian Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249–276.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 8–14. Association for Computational Linguistics.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? Assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382. Association for Computational Linguistics.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! Find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265. Association for Computational Linguistics.

Jon Sprouse and Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa*, 2(1):14.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Proccedings of the Seventh International Conference on Spoken Language Processing*, pages 901–904.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental RNNs and a syntactic scaffold. *arXiv preprint arXiv:1706.09528*.

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The importance of being recurrent for modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Julie A. Van Dyke. 2007. Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2):407–430.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Matthew W. Wagers, Ellen F. Lau, and Colin Phillips. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2):206–237.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Ming Xiang, Brian Dillon, and Colin Phillips. 2009. Illusory licensing effects across dependency types: ERP evidence. *Brain and Language*, 108(1):40–55.

Geoffrey Zweig and Christopher J. C. Burges. 2011. The Microsoft Research sentence completion challenge. Technical report, Microsoft.