An LSTM adaptation study of (un)grammaticality

Shammur Absar Chowdhury and Roberto Zamparelli
CIMeC: Center for Mind/Brain Sciences
University of Trento
{shammur.chowdhury, roberto.zamparelli}@unitn.it

Abstract

We propose a novel approach to the study of how artificial neural network perceive the distinction between *grammatical* and *ungrammatical* sentences, a crucial task in the growing field of *synthetic linguistics*. The method is based on performance measures of language models trained on corpora and fine-tuned with either grammatical or ungrammatical sentences, then applied to (different types of) grammatical or ungrammatical sentences. The results show that both in the difficult and highly symmetrical task of detecting *subject islands* and in the more open CoLA dataset, grammatical sentences give rise to better scores than ungrammatical ones, possibly because they can be better integrated within the body of linguistic structural knowledge that the language model has accumulated.

1 Introduction

As the language modeling abilities of Artificial Neural Network (ANN) expand, a growing number of studies have started to address a network's ability to distinguish sentences contain various types of syntactic errors from minimally different correct sentences, thus providing the equivalent of human *grammaticality judgments*, one of the cornerstones of theoretical linguistics since Chomsky (1957). These studies are important for at least two reasons: they can shed light on the type and amount of information which can be learned from pure linguistic data without any specialized language-learning device (thus contributing to the debate on human Universal Grammar, Chomsky 1986; Lasnik and Lidz. 2015; Chowdhury and Zamparelli 2018), and they can be used as probes on the ANNs themselves, investigating whether models which are apparently proficient at language modeling are actually sensitive to the same syntactic and semantic cues humans use.

The ANNs used in this area of research (often LSTMs, Hochreiter and Schmidhuber 1997, but recently also transformer-based ANN, Vaswani *et al.* 2017, all trained on large datasets of normal text) are tested on a mix of grammatical or ungrammatical sentences. The latter are obtained either by altering naturally occurring sentences (semi-randomly, as in Lau *et al.* 2017, or systematically, Linzen *et al.* 2016; Gulordava *et al.* 2018), by collecting examples from the published linguistic literature (Warstadt *et al.*, 2018) or by creating minimal pairs by hand (individually, Wilcox *et al.* 2018, or with sentence-schemata, as in Chowdhury and Zamparelli 2018).[1]

Once test data have been acquired, the literature has threaded between two very different approaches: treating grammaticality as a *classification* problem (i.e. feeding grammatical/-ungrammatical sentences to a classifier and asking it to discriminate, cf. the first experiment in Linzen *et al.* 2016), or feeding the test sentences to a Language Model (**LM**) pretrained on normal language and measuring the perplexity accumulated by the LM as it traverses the sentence.[2]

The classification approach works somewhat better, and can tell us if the possibility to spot un-

This work was funded by the Italian 2015 PRIN Grant "TREiL".

[1] Most studies except Lau *et al.* (2017) take the simplifying assumption that judgments can be treated as binary (e.g. *acceptable/non-acceptable*). This position is not entirely satisfactory, theoretically, but we believe that it won't do much harm at this early stage of research.

[2] Intermediate methods are possible: Warstadt *et al.* (2018) and Warstadt and Bowman (2019) train a classifier on sentence vectors produced by various types of language models.

Proceedings of the Second BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 204–212
Florence, Italy, August 1, 2019.

grammaticality can *in principle* be learned from the data, but is not directly comparable with the human ability to detect ungrammaticality, since explicit syntactic judgments play a negligible role in language acquisition.

The approach which reads (un)grammaticality from the performance of a LM starts from a more naturalistic task—predicting what's coming (van Berkum, 2010)—and can thus be more directly compared to human performances, but the probability assigned by a LM to the words reflects many factors (sentence complexity, level of embedding, semantic coherence, etc.), making it difficult to tease apart 'grammaticality' from a more general notion of 'acceptability' or 'processing load'.

In this paper we propose a third approach to measuring grammaticality, derived from the LM method. In this approach, we utilized our in-house pre-trained LSTM LM and *adapt* the model via *fine-tuning* (Pan and Yang, 2010; Li, 2012) on variations of the test sentences.

Grammaticality is then treated as a comparative measure of coherence: to what extent the new (un)grammatical input can be integrated with what the ANN has learned so far, and to what extent it can improve similar grammatical or ungrammatical constructions. We test this method with a large number of artificially generated examples, focusing on a particularly difficult contrast, the case of subject vs. object subextraction[3]. We then apply the method to a more general scenario, the CoLA dataset, tuning a LM with either *grammatical* or *ungrammatical* CoLA sentences and measuring its performance in various testing scenarios.[4]

In the following sections, we first present a detailed task description, in Section 2, followed by a brief overview of the methodology and datasets used for the study (Section 3). In Section 4, we formalize our hypothesis of how the model should behave and report the results and observation of the network behavior in Section 5; we then discuss our observation and conclude the study with future directions in Section 6.

[3]The expanded test sets for each task can be found in https://github.com/LiCo-TREiL/Computational-Ungrammaticality/tree/master/blackboxnlp2019.

[4]See Warstadt *et al.* (2018). Every sentence in the corpus, which can be found at https://nyu-mll.github.io/CoLA/, is marked as grammatical or ungrammatical. The values are drawn from the published literature, see Warstadt *et al.* (2018, Tab.2) for details.

2 Task Description

It has been noted since Ross (1967) that while Wh-questions and relatives clauses (RC) can give raise to gaps at unbounded distance (as in *Who did Mary say that John saw* __ and *The boy that Mary thinks that John adopted* __), gaps in certain positions (e.g. inside relative clauses, individual conjuncts, or certain adjuncts) are perceived as degraded. Ross coined the term *syntactic islands* for these environments, which have been the focus of a huge amount of research in theoretical linguistics (see e.g. Szabolcsi and den Dikken 1999). Studies on ANNs' sensitivity to grammaticality have tried to model certain types of islands, with varying degree of success (Lau *et al.*, 2017; Wilcox *et al.*, 2018, 2019; Jumelet and Hupkes, 2018). In this paper, we address *subject islands*, i.e. the difference between (1a) and (b) for Wh-interrogatives, and between (2a) and (b) for RCs.

(1) a. Which people did activists love [fighting for __]?
    b. *Which people did [fighting for __] appeal to activists?

(2) a. the causes that Mary feared [fighting against __]
    b. *the causes that [fighting against __] scared Mary

Subject islands are an interesting domain for various reasons: (i) extractions from subjects and object can contain nearly the same words (like above), and there are no lexical cues which signal one or the other type (e.g. both cases in (1) require *do*-support); (ii) while (1) and (2) share the extraction phenomenon, they have completely different discourse functions and distributions: is not obvious that a model that learns relative clauses should boost its processing of Wh-questions, or vice-versa; (iii) extractions out of PPs inside nominals are rare in naturally occurring data, so they stand as a challenging test of the ANN's generalization abilities.

Embedded Wh extractions out of PPs (**I know who the painting by __ fetched a high price at auction.*) were one of the violations studied in Wilcox *et al.* (2018), using Google's LM and the model from Gulordava *et al.* (2018). Neither LMs managed to model extractions out of PPs, treating the PP either as a possible extraction domain (Google's LM) or an island in both subject and

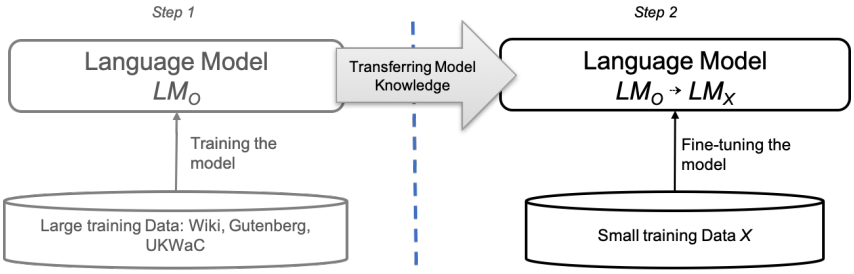

Figure 1: Experimental Pipeline.

object position (Gulorodova's). The study didn't address RCs like (2). This case thus presents an interesting challenge for our technique: combined with the sentence schemata method described in Section 3, it gives a highly controlled environment; however, this comes at the cost of a high lexical overlap (after fine-tuning, the ANN is tested on structures which contain many words it has already practiced with). To try a different and more open testing environment we applied the same method to the 5 test sets of the CoLA dataset (see Section 3 for details) . In this case, we fine-tuned the ANN on grammatical or ungrammatical sentences from the CoLA training set, and tested it on *different* CoLA test-phenomena sentences, checking the interactions. Since this part of CoLA is categorized by topic this gives a sense of which types of phenomena improve with this method.

3 Methodology

In this section, we describe our pipeline, including details of the datasets used in each steps. In addition, we present the evaluation measure used to validate the effects (if any) of the fine-tuning method on our tasks.

Figure 1 shows the pipeline we propose for exploring the effect of rehearsing new (un)grammatical input on a trained LSTM language model.

LM Architecture: The first step (Step 1 in Fig. 1) is to train a language model (LM_O) using a large text corpus. For the study, we used a left-to-right long-short term memory (LSTM) language model (Hochreiter and Schmidhuber, 1997), trained with 500 hidden units in each layer ($layers = 2$) and an embedding dimension of 256. The model was trained using a PyTorch RNN implementation with dropout regularization technique applied in different layers of the architecture, along with SGD optimizer using a fixed batch size of 80. We have not tuned the models for different dropouts or learning rate parameters, among other parameters.

Corpus	Style	%
Wiki-103	Encyclopedic data	12.15
Gutenberg[6] Dataset	Narrative style: includes collection of English books	36.58
UKWaC	Mixed, crawled from .uk domain	51.27

Table 1: Composition of the training set and style of training data.

Datasets for Training LM: To train the LSTM model, we used different English corpora — for stylistic variety — extracted from Wikipedia, the Gutenberg Dataset (Lahiri, 2014) and UKWaC (Ferraresi *et al.*, 2008), as shown in Table 1. We then tokenized the input sentences, removing URLs, email addresses, emoticons and text enclosed in any form of brackets ({.},(.), [.]). We replaced rare words (tokens with frequency < 20)[5] with *<UNK>* token along with its signatures (e.g. *-ed, -ing, -ly* etc.) to represent every possible out-of-vocabulary (OOV) words. We also replaced numbers (exponential, comma separated etc) with a *<NUM>* tag. We removed the sentences from UKWaC with OOV tags. Therefore, to train LM we used a training set consisted of $\approx 0.7B$ words in $\approx 31M$ sentences, with a vocabulary of size $|V| = 0.1M$.

Adaptation via fine-tuning: The trained LM_0 was used to initialize the weights of the new LSTM LM_X, so as to transfer the knowledge

[5]For preparing the vocabulary set V, we only considered tokens presents in WikiText and Gutenberg dataset.

[6]We intentionally removed the stories that overlapped with the test and dev set of Childrens Book Test (CBT) (Hill *et al.*, 2015), for training purpose.

LM_0 has acquired so far (Step 2 Fig. 1). To adapt the models LM_X to new (un)grammatical structures, we fine-tuned the models by feeding the sentences from our small training data sets, with batch size of 20 and epoch e ($e = \{3, 10\}$). All other parameters remained unchanged with respect to the original LM_0. In this paper, for brevity, we only report the results after 3 epochs.

Datasets for Adaptation: LMs can be quite sensitive to the specific content words used. To minimize this effect and focus on structure, we used the 'sentence schemata' method from Chowdhury and Zamparelli (2018): starting from a schema such as (3), a script automatically generates sentences containing all the possible combinations of the bracketed expressions. The schema in (3) (tagged **Aff**(irmatives with complex) **Obj**(ects)) gives 160 affirmative sentences (e.g. *Activists hated fighting for these laws*); we also constructed schemata for affirmatives with the gerund in subject position (**AffSubj**, e.g. *fighting for these causes scares politicians*), as well as for the corresponding root Wh-clauses (**WhSubj**, **WhObj**, as in (1a)/(1b), and relatives (**RelSubj, RelObj**, as in (2a)/(2b)).[7] In total, we have 6 train/test sets, see Table 2 for details.

(3) [John Mary politicians activists governments] [feared loved hated thought_about] fighting [for against] these [causes movements people laws] .

Apart from exploring adaptation of subject islands, we also explored the effect of adaptation in an open testing environment (as mentioned in Section 2). For this setting, our training and testing data is less likely to have a substantial lexical overlap. For the adaptation part, we split the CoLA training set in two parts—one consisting of grammatical sentences ($CoLA_G$), the other one of ungrammatical sentences ($CoLA_{UG}$), both covering different linguistic phenomena such as islands, passives, coordination, negative polarity, etc. As test sets, we used different CoLA-test phenomena.[8] They are:

- Subject-Verb-Object (**SVO**): The test set consists of utterances, generated using different permutation of subject (S), verb (V) and object (O). The set containes 10 subjects, 2 verbs and 5 objects.

- Wh-Extraction (**WhExt**): This set tests the ability to note that a Wh- must correspond to a gap, with pairs such as *What did John fry? / *What did John fry the potato?* (cf. Wilcox *et al.* 2018, Sec.2.3, Chowdhury and Zamparelli 2018, Task B).

- Causative-Inchoative Alternation (**CausAlt**): Based on verbs that do or do not undergo the alternation (*Kelly popped/blew the bubble.* vs. *The bubble popped/*blew.*).

- Subject-Verb Agreement (**SVAgr**): A set based on number agreement mismatch, such as *the child (that was accompanied by his parents) has/*have left*. This is the task used in Linzen *et al.* (2016); Gulordava *et al.* (2018).

- Reflexive-Antecedent Agreement (**ReflAgr**): A test on whether reflexive pronouns have appropriate local antecedents (cf. *I amused myself / *yourself / *herself / *him- self / *ourselves / *themselves*).

Evaluation Measure: To track the performance of our LSTM on the test sets, we adopted the popular acceptability measure *Syntactic log-odds ratio* ($SLOR$), introduced in this domain by Lau *et al.* (2017) and shown in Equation 1.

$$SLOR(\varepsilon) = \frac{log(p_m(\varepsilon)) - log(p_u(\varepsilon))}{|\varepsilon|} \quad (1)$$

where ε represents the sentence; $p_m(.)$ is the probability of the ε given by the model, calculated by multiplying probabilities of each target words, present in the sentence; $p_u(.)$ is the unigram probability of the ε and $|\varepsilon|$ represent the length of the sentence.

The measure considers the structure and position of the words, subtracting out the unigram log-probability so that sentences that use rare words are not penalized, and is normalized by sentence length, thus removing (positive or negative) biases due to long sentences. Higher SLOR values correspond to 'better' (i.e. more predictable/acceptable) sentences.

[7] For training, we merged AffObj and AffSubj to create a general set of affirmatives, Aff.

[8] Please check Warstadt *et al.* (2018, Tab. 2) for details.

Sets	# inst.	Used for	Sets	# inst.	Used for
Close Environment Testing			Open Environment Testing		
AffObj: *X likes* [*fighting for Y*]	160	test	CoLA_G	6029	train
AffSubj: [*fighting for Y*] *pleases X*	160	test	CoLA_{UG}	2532	train
Aff: AffSubj∪AffObj	320	train	SVO	500	test
RelObj: *the Y that X likes* [*fighting for* __]	160	train/test	WhExt	520	test
RelSubj: *the Y that* [*fighting for* __] *pleases X*	120	train/test	CausAlt	182	test
WhObj: *What did X like* [*fighting for* __]*?*	200	train/test	SVAgr	676	test
WhSubj: *What did* [*fighting for* __] *please X ?*	150	train/test	RiflAgr	144	test

Table 2: Detailed information of the training/testing set for the adaptation experiments. inst. represent instances

4 Our Hypothesis

We expect the adapted LM to improve in proportion to the similarity between the tasks, but also in proportion to how well the material presented in the fine-tuning learning phase is consistent with what the ANN already knows about language structures.

Our expectations are that retraining with ungrammatical sentences should be harder to incorporate into previous knowledge, thus leading to worse performances in terms of generalization. Note that improvements when the ANN is trained on Wh and tested on RC or vice-versa can be attributed in part to lexical familiarity (the training contained most of the words seen in the testing), in part to the model's ability to note the common element in the two constructions, i.e. the extraction. We can mitigate the lexical overlap problem by subtracting the scores of a LM fine-tuned on the affirmative cases (i.e. the sentences generated from (3)) from those obtained from the corresponding extraction cases (RC and Wh), since our affirmative cases already contain most of the lexicon found in the RC/Wh sentences.

In the second experiment, where we tested on CoLA, there is no reason to expect a very high lexical overlap, so any effect found there can be attributed purely to the structures.

5 Results

Subject/Object Extraction Figure 2 gives an overview of the SLOR values of our LSTM tuned for 3 epochs just on the affirmative sentences (LM_X, left), compared to the original (LM_O, right). Unsurprisingly, the LM_X shows a large improvement in the AffSubj/AffObj cases, but also an improvement in Wh case and especially in relative clauses. Note that after fine-tuning, all conditions (Aff,Rel and Wh) show a significant preference for the object case (present in Aff/Wh even in the original run). This effect emerged also in Chowdhury and Zamparelli (2018) (and in work of ours, under review, which specifically addresses this phenomenon). Since it is also present in affirmatives, it cannot obviously be attributed to a sensitivity to islands, but can probably be put down to a general preference of LSTM LMs for having complex structures in object position. This effect seems to overcome an effect found in Chowdhury and Zamparelli (2018) (Task A, which however uses different measures), where subject relatives scored better than object relatives (while both being grammatical), in line with human parsing preferences widely discussed in the psycholinguistic literature (Gibson, 1998; Gordon *et al.*, 2001; Friedmann *et al.*, 2009). The general lower score for RCs, compared to Wh cases, could also be attributed to the fact that in the testing phase the LM receives an End-of-Input signal before the sentence is over (i.e. RC are sentence fragments).

Figure 3 shows the effect of fine-tuning the original LM on different parts of the test set and testing it on the others. At a global level, if we compare the scores with the affirmative baseline (the performance of the model fine-tuned with affirmatives only, as in Figure 2, left), we see that on average adding Wh-clauses significantly boosts RCs (+0.64) and vice-versa, though not as strongly (+0.39). Next, tuning with *grammatical* material gives a larger overall boost than tuning on *ungrammatical* material. This can be seen from the Total in Table 3 (using the notation ARelObj(RelObj-Aff(RelObj)) to mean "SLOR of LM_O fine-tuned with Aff+RelObj (ARelObj) and applied to Relatives with OBJect subextraction minus the SLOR of the test set using model adapted by Aff)"). Within construction, tuning with Aff plus Obj extractions boosts other object cases (green cells)

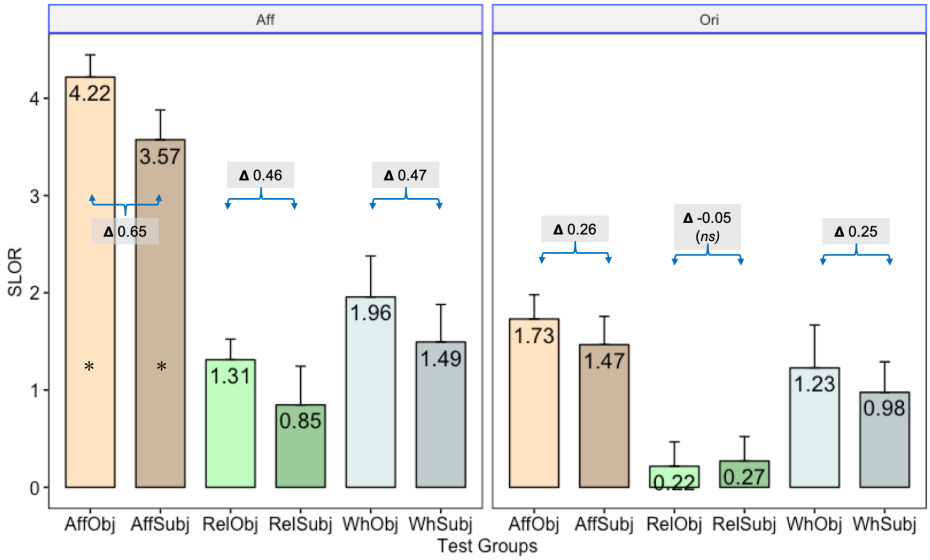

Figure 2: Variation of SLOR measure for different test groups using the model adapted on affirmative sentences (Aff, both AffSubj and AffObj) and the original LM_0 (Ori) model. Higher is better. The blue arrows with the Δ values represents the difference in SLOR between grammatical and ungrammatical sentences. * warns that the same testset is used to adapt the respective model. *ns* indicates that the results are not significantly different from each other.

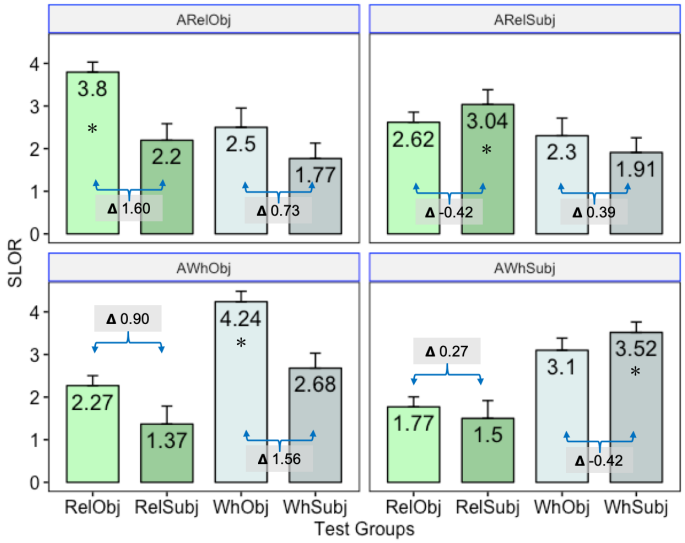

Figure 3: Variation of SLOR measure for different test groups using models adapted on: relative clause-object (ARelObj); relative clause-subject (ARelSubj); wh-object (AWhObj); wh-subject (AWhSubj). All the models are initially adapted on affirmative sentences, hence the presence of A in **A**RelObj and all other models. The blue arrows with the Δ values represents the difference between the SLOR of grammatical correct sentences with ungrammatical sentences. The * warns when the same testset was used to adapt the corresponding model.

more than tuning on Aff plus Subj extractions boosts other Subj cases (pink cells); across construction, Aff+WhObj tuning boosts RelObj and even RelSub and, to a lesser extent, Aff+RelObj tuning boosts WhObj more than Aff+RelSubj boosts WhSubj.

CoLA results Figure 4 shows the results on the 5 test sets for the original model (LM_O) and the LM fine-tuned with the CoLA grammatical and ungrammatical sentences, respectively. The first thing to note is that LM_O is already able to significantly distinguish, on average, the two classes, with the worst performances coming from the Causative-Inchoative Alternation, a construc-

		Fine-tuned with			
	Testing scores	ARelObj	ARelSubj	AWhObj	AWhSubj
a.	**RelSubj–Aff(RelSubj)**	1.35	2.19	0.52	0.65
b.	**RelObj–Aff(RelObj)**	2.49	1.31	0.96	0.46
c.	**WhSubj–Aff(WhSubj)**	0.28	0.42	1.19	2.03
d.	**WhObj–Aff(WhObj)**	0.54	0.34	2.28	1.14
e.	**Total**:	4.66	4.26	4.95	4.28

Table 3: Effect of fine-tuning. A**X** represent the models tuned with affirmatives followed by **X**. **Y**-Aff(**Y**) represent the test scores (SLOR) using the particular model minus the SLOR of the model adapted on affirmatives (Aff) for the Y test set. $X, Y \in \{RelObj, RelSubj, WhObj, WhSubj\}$.

tion linked to the lexical semantics of a class of verbs which are not likely to be encountered in many other examples. As in the previous experiment, fine-tuning improves the SLOR scores of all cases, ungrammatical ones included. In keeping with the previous experiment, we verify whether the switch from CoLA$_G$ to CoLA$_{UG}$ has a significant effect on the improvements (esp. CoLA-G(G) vs. CoLA-UG(UG), keeping in mind that here, unlike in the previous experiment, the training can contain at most a small dose of the lexicon and the phenomena in the testing set). Given the results in the Subj/Obj island task, our expectations are that tuning on CoLA$_G$ should work better than tuning on CoLA$_{UG}$. The difference turns out not to be significant with Subject-Verb Agreement cases (SVArg, Figure 4a), significant but with ungrammatical cases coming out best for the Subject-Verb-Object permutation cases (SVO, Figure 4b), significantly bigger with grammatical tuning in the remaining cases (see 4f for the overall picture). The case of SVArg might be due to the fact that the contrastive examples found in the syntactic literature might not cover something as basic as wrong subject-verb agreement. The behavior of SVO remains unclear.

6 Discussions and Conclusions

The results of ours first experiment suggest that, even though the contrast between subject and object subextraction is one of the hardest for ANNs to detect (see Wilcox *et al.* 2018), fine-tuning a language model with one of the two conditions does not give the same effect: above and beyond the effect of assertions (see Figure 3), tuning with grammatical extractions (i.e. object cases) yields a larger boost for the construction used for tuning than tuning with the ungrammatical cases. In small measure, the boost extends to the related construction (Wh to RC, and partially vice-versa). The same effect is found with the much less controlled CoLA dataset, at least for some of the constructions we tested.

The results are consistent with the hypothesis that grammatical cases are somehow easier to integrate into what the ANN has already discovered about linguistic structures. Of course, positive examples of grammatical extractions like WhObj and RelOBj also boost the ungrammatical cases, but possibly this is because they apply to parts of the sentence different from the extraction site (indeed, ungrammatical cases boost grammatical and ungrammatical cases almost to the same degree). This suggests that the methodology we are proposing could be a useful addition to the toolbox of this research area.

An obvious question, at this point, is whether the fine-tuning approach could be turned into a *classification* method. One could for instance imagine classifying a sentence as grammatical or ungrammatical on the basis of its SLOR difference across LMs tuned with grammatical/ungrammatical sets (e.g. CoLA$_G$ and CoLA$_{UG}$ conditions). Recall however that SLOR is sensitive to a variety of factors which have nothing to do with grammaticality (e.g. collocations, pragmatic plausibility), and that it has been used to study grammaticality only with carefully constructed minimal pairs. While not impossible, we suspect that a classification experiment could not be done with relatively open data like CoLA, though it is possible that with more balanced materials such an experiment might become possible. Probably a better use for the technique proposed here would be to study similarity across constructions *as seen by the network*. Using the more fine-grained classification of the CoLA data given in Warstadt and Bowman (2019), it might be possible to selectively fine-tune a model with one construction, test it with all the others and discover

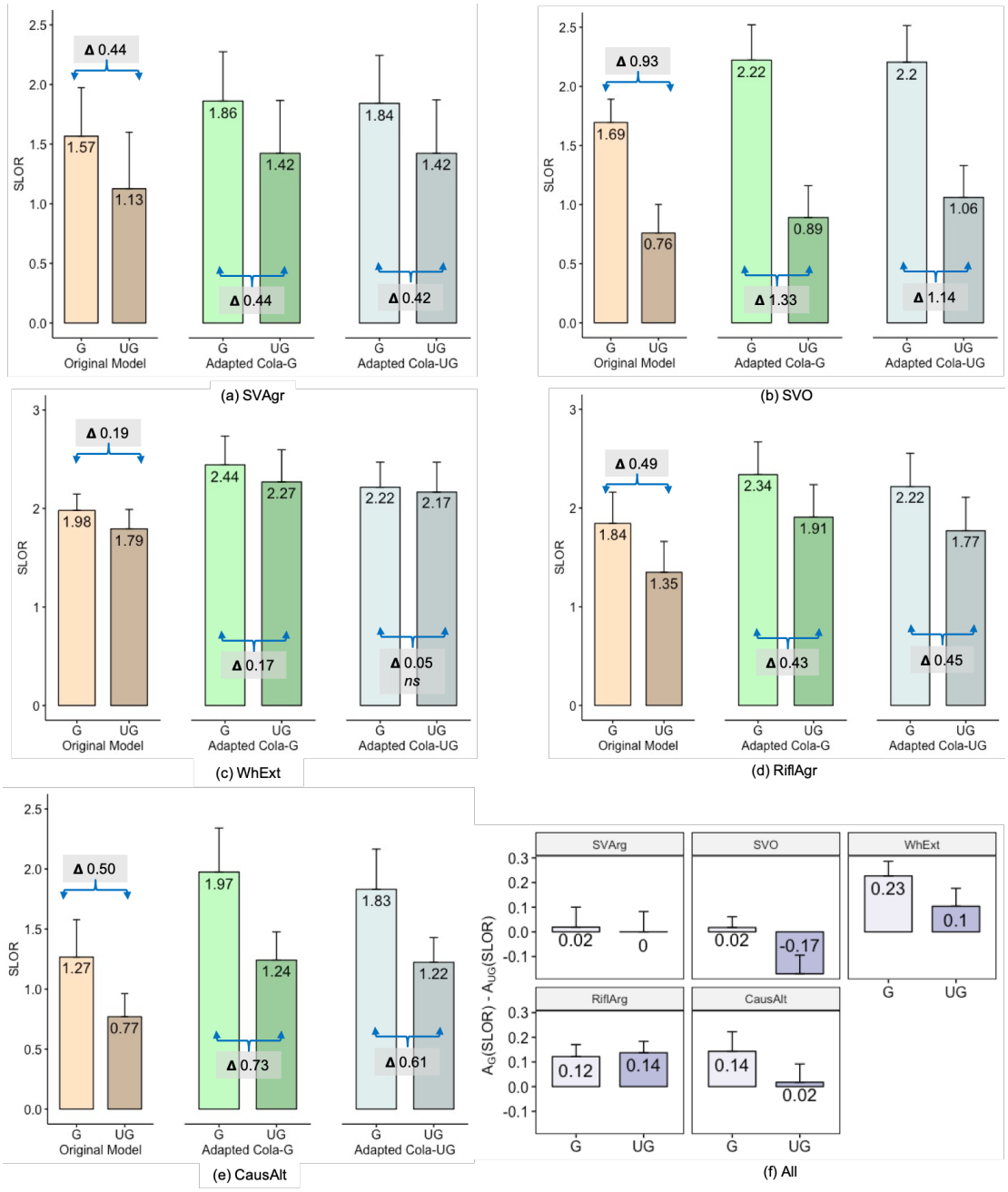

Figure 4: Variation of global SLOR measure for: Figure 4a - Subject-Verb Agreement (SVAgr); In Figure 4b - Subject-Verb-Object (SVO); In Figure 4c - Wh-Extraction (WhExt); In Figure 4d - Reflexive-Antecedent Agreement (RiflAgr); Figure 4e - Causative-Inchoative Alternation (CausAlt). For all Figure 4 a-e, Original model is un-adapted LM model, where as Adapted Cola-G(UG) represent the results from the model which is adapted on CoLA train grammatical (ungrammatical) instances. Figure 4f represents the difference in the measure of SLOR value, for grammatical (G) and ungrammatical (UG) examples, between Cola-G and Cola-UG model, i.e. $A_G(SLOR) - A_{UG}(SLOR)$ for all the above test cases (a-e), where A_G represents result from Cola-G model and similarly A_{UG} represents result from Cola-UG. The blue arrows with the Δ values represents the difference between the SLOR of grammatical correct sentences with ungrammatical sentences. *ns* indicates that the results are not significantly different from each other.

from the variations in a performance measure like SLOR how the ANN 'sees' the relation between different linguistic cases.

References

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origins and Use*. Praeger, New York.

Chowdhury, S. A. and Zamparelli, R. (2018). RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144. Association for Computational Linguistics.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop WAC4*, pages 47–54.

Friedmann, N., Belletti, A., and Rizzi, L. (2009). Relativized relatives: types of intervention in the acquisition of a-bar dependencies. *Lingua*, **119**(1), 67–88.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, **68**, 1–76.

Gordon, P., Randall, H., and Marcus, J. (2001). Memory interference during language processing. *J. Exp. Psychol. Learn. Mem. Cogn.*, **27**, 1411–1423.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL*.

Hill, F., Bordes, A., Chopra, S., and Weston, J. (2015). The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

Jumelet, J. and Hupkes, D. (2018). Do language models understand anything? on the ability of lstms to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium.

Lahiri, S. (2014). Complexity of Word Collocation Networks: A Preliminary Structural Analysis. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 96–105, Gothenburg, Sweden. Association for Computational Linguistics.

Lasnik, H. and Lidz., J. (2015). The argument from the poverty of the stimulus. In I. Roberts, editor, *Oxford Handbook of Universal Grammar*, pages 221–248. Oxford University Press.

Lau, J. H., Clark, A., and Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, **41**(5), 1202–1241.

Li, Q. (2012). Literature survey: domain adaptation algorithms for natural language processing. *Department of Computer Science The Graduate Center, The City University of New York*, pages 8–10.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 521–535.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, **22**(10), 1345–1359.

Ross, J. R. (1967). *Constraints on Variables in Syntax*. Indiana University Linguistics Club, Bloomington.

Szabolcsi, A. and den Dikken, M. (1999). Islands. *GLOT Internationaal*, **4**(6), 3–8.

van Berkum, J. J. A. (2010). The brain is a prediction machine that cares about good and bad – any implications for neuropragmatics? *Italian Journal of Linguistics*, **22**.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Warstadt, A. and Bowman, S. R. (2019). Grammatical analysis of pretrained sentence encoders with acceptability judgments.

Warstadt, A., Singh, A., and Bowman, S. R. (2018). Neural network acceptability judgments. https://arxiv.org/abs/1805.12471.

Wilcox, E., Levy, R., Morita, T., and Futrell, R. (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, Brussels. ACL.

Wilcox, E., Qian, P., Futrell, R., Ballesteros, M., and Levy, R. (2019). Structural supervision improves learning of non-local grammatical dependencies.