Priorless Recurrent Networks Learn Curiously

Jeff Mitchell
School of Psychological Science
University of Bristol
Bristol, UK
jeff.mitchell@bristol.ac.uk

Jeffrey S. Bowers
School of Psychological Science
University of Bristol
Bristol, UK
j.bowers@bristol.ac.uk

Abstract

Recently, domain-general recurrent neural networks, without explicit linguistic inductive biases, have been shown to successfully reproduce a range of human language behaviours, such as accurately predicting number agreement between nouns and verbs. We show that such networks will also learn number agreement within unnatural sentence structures, i.e. structures that are not found within any natural languages and which humans struggle to process. These results suggest that the models are learning from their input in a manner that is substantially different from human language acquisition, and we undertake an analysis of how the learned knowledge is stored in the weights of the network. We find that while the model has an effective understanding of singular versus plural for individual sentences, there is a lack of a unified concept of number agreement connecting these processes across the full range of inputs. Moreover, the weights handling natural and unnatural structures overlap substantially, in a way that underlines the non-human-like nature of the knowledge learned by the network.

1 Introduction

Grammarians have for a long time (e.g.: Pāṇini, 6th Century BCE; Aristotle, 350 BCE) been interested in characterising what is and is not possible within a given specific language. Famously, Chomsky (1957) used the sentence *colorless green ideas sleep furiously* and its reversal *furiously sleep ideas green colorless* to illustrate the difference between grammatical and ungrammatical within English. In addition, modern linguistics has also been concerned more generally with what is and is not possible within the space of all human languages.

In this article, we investigate the ability of a neural language model to learn structures on the impossible side of this boundary. That is, we train and test the network on data which has been manipulated to contain unnatural structures. In contrast, the evaluation of such models has usually focused on the other side of the boundary, i.e. on data gathered from real human behaviour within specific natural languages. Here, we are interested in the more general question of whether these neural architectures are appropriate models of human language capacities, in terms of what differentiates possible from impossible structures.

From a theoretical perspective, a number of grammatical formalisms have been proposed as computational models of the space of possible natural languages: from Context-Free Grammars (Chomsky, 1956) to Tree-Adjoining Grammars (Joshi et al., 1969) and Combinatory Categorial Grammars (Steedman, 1987). One characteristic that all such grammars share is the organisation of linguistic expressions into hierarchical, as opposed to purely linear, structures.

Empirical research has also uncovered further design patterns common across many superficially diverse languages. Although the concept of language universals is not uncontroversial, most linguist would agree that, at the very least, natural languages display strong shared tendencies and similarities. For example, languages typically have a preferred word order - e.g. Subject Verb Object vs. Subject Object Verb - which is correlated with the ordering of other constructions - e.g. prepositions vs. postpositions.

Proceedings of the 28th International Conference on Computational Linguistics, pages 5147–5158
Barcelona, Spain (Online), December 8-13, 2020

Alongside the observation that children learn their native languages efficiently from limited, noisy input, these shared principles have been been used to support the view that the human language faculties are innately adapted to learn and process one specific set of possible languages. Or, to put it another way, there are structures that are essentially impossible for these innate faculties to handle.

Recently, however, domain-general neural sequence models, lacking explicit linguistic inductive biases, have demonstrated impressive performance on a range of practical language processing applications. This, in turn, has led to increasing interest in the question of how effective these architectures are as models of general human linguistic behaviour. Whereas our interest here is in their behaviour when trained on unnatural structures, most research has so far been interested in their ability to successfully learn the structure of natural languages.

For example, Gulordava et al. (2018) examined the extent to which a Long Short Term Memory (Hochreiter and Schmidhuber, 1997) network trained on raw text learned about agreement phenomena across a number of languages. They found that, given the right training data, the model was able to predict long-distance agreement on both natural and semantically nonsensical but syntactically valid sentences. Similarly, Futrell et al. (2019) investigated the ability of such models to track syntactic state within subordinations and garden path sentences. Treating their networks as psycholinguistic subjects, they found that given enough data the models responded to subtle cues in a manner comparable to human subjects.

These results suggest that although an LSTM handles its input in a purely sequential manner, it is able, nonetheless, to learn about phenomena that have traditionally been represented within hierarchical data structures. Frank and colleagues (Frank, 2009; Frank and Bod, 2011; Frank et al., 2012; Frank et al., 2015) suggest that, in fact, such sequential models have advantages over hierarchical alternatives as cognitive models of language processing. Using evidence from reading time and ERP studies, they argue that the ability to accurately model the probabilities of words in context is more important in a psycholinguistic model than the particular architecture employed.

Given these successes in language learning, Pater (2019) has proposed a complementary role for neural networks alongside generative linguistics. He argues that these architectures may supply the theory of learning that linguistics currently lacks. In response, Rawski and Heinz (2019) invoke the no-free-lunch theorems (Wolpert and Macready, 1997) and poverty-of-the-stimulus arguments (Chomsky, 1986) to question whether neural models actually have the right inductive biases.

Here, we investigate this question by evaluating a network's ability to learn number agreement within pathological structures unlike any natural language. While failure to learn natural structures would give important insights into a model's limitations, success on unnatural structures is also revealing, suggesting a mismatch between the model's architecture and human linguistic processes. We therefore train and test an LSTM on manipulated English sentences and compare this performance to that on the original unmanipulated corpus. Based on prior research using unnatural data, we employ three types of modification: reversing word order, inserting a linear sequence of tokens, and shuffling the vocabulary.

Introducing reversed constructions alongside the original sequences disrupts a language's preference for a consistent word order and so is an effective way of producing unnatural data. Moreover, the human experiments described in Section 2 suggest that learners struggle to acquire artificial languages constructed in this way. Learners also struggle with constructions based on sequential counting, e.g. inserting a word in the nth position regardless of the syntactic context. Human languages universally base their syntax around hierarchical structures, and so constructions based on counting and linear order are entirely unnatural.

While word order preferences and hierarchical structure are language specific characteristics differentiating natural from unnatural data, machine learning theorists have also proposed domain general methods for generating unnatural data. In this regard, shuffling provides a useful means for for analysing the limits of learning algorithms. In particular, fully randomised shuffling of the training labels in a supervised learning task forces an algorithm to resort to memorisation of the training data, supplying insight into its capacity. In our experiments, we employ a limited form of shuffling, where generalisation is still possible, but which increases the computational load on the model substantially.

We describe these experiments and their results in Section 3. Following that we analyse in more detail how knowledge about number agreement is encoded in the weights of the network in Section 4. First, however, we describe some of the prior work on unnatural data in Section 2.

2 Learning and Unnatural Data

Performance on unnatural data has been discussed by a number of researchers interested in understanding the characteristics of learning processes in humans and machines. Such a perspective can provide insight not only into the limits of human learning, but also into the lack of such limits in machines. For example, Pinker and Prince (1988) critique a connectionist model of past tense formation in terms of its handling of unnatural structures. They argue that the neural architecture proposed by Rumelhart and McClelland (1986) cannot be an adequate model of learning because it would permit string reversal as a means of forming the past tense, which is never attested in natural languages.

Smith et al. (1993), in contrast, studied the learning of unnatural linguistic structures by humans rather than machines. Working with a polyglot savant, Christopher, and a group of normal adult controls, they investigated second language learning of an artifical language, Epun, containing both natural and unnatural grammatical structures. They found that while both Christopher and the controls could master the linguistically natural aspects, only the controls could eventually handle the structure dependent unnatural phenomena, such as reversing subject and verb to express negation, and neither could master the structure independent aspects, such as marking emphasis on the third word of a sentence. They argue that Christopher's abilities are entirely due to his intact linguistic faculties, but that the controls had recourse to employing other general cognitive abilities for some phenomena.

This proposal, that second language learners may employ distinct cognitive resources in handling natural and unnatural structures, is supported by the fMRI evidence of Musso et al. (2003). In that experiment, native speakers of German were taught modified versions of Italian and Japanese, containing unnatural constructions alongside their real grammar. For example, in both languages, interrogatives were constructed by reversing a declarative sentence and the marker for negation was placed after the third word. While the learners mastered both the natural and unnatural phenomena, reaction times were faster for natural structures. Moreover, the brain imaging revealed that activation of Broca's area was associated with learning of the natural structures, whereas performance on the unnatural phenomena was negatively correlated with this activation.

Addressing the question of whether statistical methods can provide cognitively faithful models of human language knowledge, Fong and Berwick (2008) investigated the performance of a statistical parser (Bikel, 2004; Collins, 1996) when trained and tested on unnatural structures. They found that the accuracy of the parser remained largely unchanged on a corpus in which the orders of verbs and their complements and/or arguments and adjuncts had been reversed in every second sentence. Given this insensitivity to the consistency of word order in its training corpus, Fong and Berwick (2008) conclude that the parser's performance points to a non-human-like ability.

Extending this line of enquiry, (Fong et al., 2013) evaluated the ability of a parser to handle the impossible interrogatives studied by Musso et al. (2003). They found that both the Stanford unlexicalised parser (Klein and Manning, 2003) and Collin's parser (Bikel, 2004; Collins, 1996) were able to learn the reversed questions reasonably successfully, particularly when given enough training examples. Again, this success in learning from a corpus without consistent word order preferences was taken as evidence for a failure in modelling human language knowledge.

The learning of unnatural structures has also been investigated within neural image recognition. Zhang et al. (2017) show that state-of-the-art CNNs will easily fit to a random labeling of the training data, even when the true images are replaced with unstructured noise. These experiments were inspired by theoretical connections between generalisation ability and Rademacher Complexity (Bartlett and Mendelson, 2003), which measures the capacity of a model in terms of its ability to fit randomly switched labels. Further investigations (Belkin et al., 2019; Nakkiran et al., 2020) suggest that satisfactory in-domain generalisation can frequently be achieved by pushing the number of parameters well beyond the point at which such rote memorisation becomes possible, and allowing the network to smoothly interpolate

between training examples. This does raise questions, however, about the performance of such an interpolation strategy on out-of-domain generalisation in comparison to human judgements, particularly on adversarial and fooling images (Dujmović et al., 2020).

3 Number Agreement in Unnatural Language Structures

We apply the LSTM architecture of Gulordava et al. (2018) to modified versions of their original datasets. Based on the prior work described in Section 2, three types of modification are used: reversing the order of tokens, inserting a linear sequence of tokens and shuffling the vocabulary. In each case, we train and test the LSTM on the modified data, evaluating the ability of the model to handle number agreement within structures that deviate substantially from natural constructions.

In the first modification, we choose a random point within the sentence, insert a special marker, and reverse the remainder of the sentence beyond that point. So, for example, *The man whose dog barks likes apples* becomes *The man whose dog* $\langle marker \rangle$ *apples likes barks*. Reversal has been a common manipulation in the generation of unnatural structures (Smith et al., 1993; Musso et al., 2003; Fong et al., 2013; Fong and Berwick, 2008) because it disrupts the tendency for languages to have a preferred word order. Our transformation introduces both a strong contextual influence on word order and also highly tangled dependencies involving constituents that cross the boundary. Thus, linear position within the surface form of the sentence in relation to the marker determines, for example, whether adjectives follow nouns or precede them, whether prepositions or postpositions are used, or whether an SVO or OVS structure is followed. Moreover, constituents that cross the boundary, e.g. *whose dog barks*, become discontinuous and their dependencies become entangled with those of the intervening material. We evaluate the ability of the model to handle dependencies that cross this boundary, by placing the marker immediately before the target verb at test time.

In the second transformation, we again choose a random point in the sentence and insert a marker. In this case however, a random vocabulary item is then selected and repeated 23 times before the sentence resumes. Such counting-based linear structures have also been used in prior research (Smith et al., 1993; Musso et al., 2003) as a contrast to the hierarchical structures usually found in natural languages. The repetition of the same token introduces a construction without any of the internal structure, such as dependencies between heads, modifiers and complements, that characterises the syntactic trees found in natural data, and is instead just a linear sequence of the appropriate length. Moreover, at test time we use the incorrect form of the verb as the inserted item, placing it just before the correct form. Thus, *The dog barks* becomes *The dog* $\langle marker \rangle$ *bark bark . . . bark barks*. Correctly predicting number agreement between *dog* and *barks* requires counting the intervening *bark* tokens and accurately anticipating where the sentence will resume.

The third transformation leaves the syntactic structure of the sentence unchanged, and instead permutes the vocabulary. This is inspired by the shuffling of image labels used by Zhang et al. (2017). We first take a list of the whole lexicon, randomly shuffle it, and then use this to define a mapping from each vocabulary item to another. A sentence in the training corpus is then transformed by selecting a random point, placing a marker, and replacing all the tokens in the remainder of the sentence with their shuffled counterparts. Thus, *The man whose dog barks likes apples* becomes *The man whose dog* $\langle marker \rangle$ *and copy Simpson*. At test time, we evaluate the ability of the model to predict number agreement across this boundary by placing the marker immediately before the target verb. Agreement is then measured in terms of the model's ability to predict the shuffled vocabulary item corresponding to the correct form of the original verb.

Our experiments are based on the model and datasets of Gulordava et al. (2018). We retrain their language model using the original hyperparameters, after applying the transformations described above to their training data, consisting of ∼90M words from English Wikipedia, which they filtered to exclude sentences containing more than 5% unknown words outside a 50K vocabulary. The model was a two layer LSTM, with 650 units in the input embeddings and each recurrent layer, trained using the standard language modelling objective of predicting the next word.

After 40 epochs of training, we applied this model to their number agreement test corpus, and evaluated

Dataset	Accuracy	Std. Err.
Original	79.3%	± 2.5%
Reversed	79.3%	± 1.6%
Repeated	73.2%	± 4.6%
Shuffled	75.6%	± 1.7%

Table 1: Mean number agreement accuracies across 4 training runs on the original and modified datasets.

the model's ability to predict correctly whether a singular or plural verb is appropriate in each sentence. In particular, the LSTM is given the initial words preceding the verb, and the logits corresponding to the singular and plural forms for the next word prediction are compared. The model's outputs are evaluated as correct whenever the form attested in the sentence is given the higher value, and we report percentage accuracy averaged over 5 training runs. Gulordava et al. (2018) extracted these evaluation sentences from the English Universal Dependency treebank (Nivre et al., 2016), identifying constructions in which a verb target, morphologically marked for number, is preceded by its noun subject cue, which also agrees in number.

3.1 Results

Performance on the number agreement tasks, as detailed in Table 1, does not appear to show substantial changes in accuracy across the various dataset modifications, and under pairwise t-tests none of the differences is significant. Of the two structural transformations - reversing the end of the sentence and inserting a repeated token - the latter produces the greatest drop in performance, even though it is, to some extent, the simpler of the two. Whereas dealing with the inserted tokens is merely a question of counting accurately, the reversal transformation requires untangling complex crossing dependencies. Nonetheless, the LSTM learns to handle these structures as effectively as ordinary English.

Shuffling the vocabulary has an intermediate effect on performance despite the fact that this alteration, in effect, doubles the work that each token embedding has to do. Each token behaves in two fairly distinct ways depending on whether it occurs before or after the marker and the embeddings have to encompass both uses. Moreover, not only does the model have to handle the dependencies within the original and shuffled vocabularies, it also has to deal with the dependencies that cross the boundary between the two.

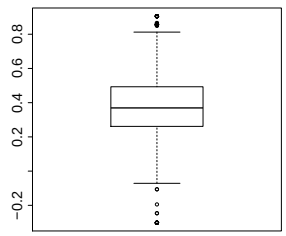
3.2 Discussion

The relative success on learning the unnatural structures in the modified datasets suggests that the network lacks biases towards the forms of grammar observed in human languages. In particular, the ability to handle extended repetitions of a randomly selected token indicates that these architectures are not biased towards structures that exhibit the sort of hierarchical phrase structure found in natural languages. Instead, they are largely indifferent to the lack of such syntactic dependencies both internally within the block of repetitions and also externally to the rest of the sentence, that this modification introduces.

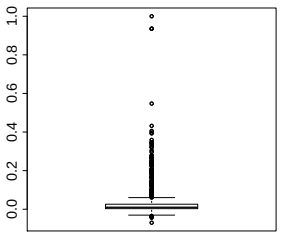
Equally, the tangled dependencies and inconsistent word orders, that reversing the end of a sentence introduces, are not an obstruction to learning. Thus, not only is no syntax an acceptable structure for the model, but these pathological dependencies are also. In addition to indicating a lack of bias towards human language structures, the success in learning the reversed structures also indicates substantial spare capacity in the original model to allow the learning of both normal and reversed word orders in the same set of weights.

This spare capacity is also indicated by the success on learning the shuffled sentences. Beyond the unnatural radical ambiguity it imposes across the lexicon, the shuffling manipulation also substantially increases the amount of information to be encoded in the network weights. In terms of number agreement, singular and plural features have to be learned in both the original and shuffled vocabularies.

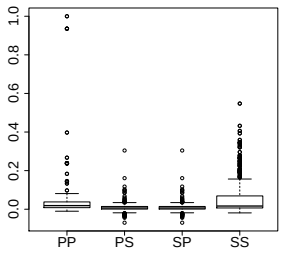
These considerations raise the question of how information about structure and features is encoded in the network's weights. In the following section, we use representations based on gradient vectors in the



(a) Similarities between different target verbs on the same prefix.



(b) Similarities of complete evaluation instances, i.e. different prefix and different target.



(c) Instance similarities broken down by correct form of each sentence (S: singular, P: plural).

Figure 1: Similarities calculated on the original model applied to original evaluation sentences. Each data point corresponds to a comparison between weight vectors associated with a pair of number agreement predictions.

space of weights to analyse the model's knowledge of number agreement.

4 Analysis of Internal Model Structure

Lakretz et al. (2019) attempted to understand the syntactic mechanisms learned by the LSTM of Gulordava et al. (2018) in terms of identifying specialised units that encoded number agreement and by tracking their activity through time. This approach uncovered long distance agreement units that retained the number information from a noun subject across intervening clauses until arriving at the verb. While this provides important insights into the kinds of processes a model learns to apply to syntactic phenomena, the analysis uncovered just two such specialised units, and their overall contribution to performance was limited. Moreover, this approach does not directly enable us to compare pairs of sentences or to investigate processes of generalisation between them.

In this article, we are instead interested in the relation between natural and unnatural sentences, and the way that the model learns their structures. As a consequence, we take a slightly different approach here, based on comparing which weights are important to predicting number agreement correctly on pairs of sentences. The weights, unlike the activities, remain fixed throughout the processing of a sentence, and are also the locus of learning and generalisation. Thus, this approach allows us to construct a single representation for a given number agreement prediction that summarises which parts of the network were critical in processing the entire preceding sentence context.

Across the majority of neural architectures, backpropagation is the standard technique of credit assignment for deciding which weights are contributing most to the current objective or loss, and as a consequence how they need to be changed. In our case, backpropagation through time allows us to compute the gradients in weight space of any objective of interest, defined in terms of the LSTM's outputs, given a sentence prefix input. Such gradients essentially tell us which weights were important in producing that value, in the sense of changes in more important weights producing larger changes in the objective. Inputs with highly similar gradient vectors therefore share a common set of important weights, or, in other words, are the product of overlapping parts of the architecture. In contrast, two inputs with orthogonal gradient vectors rely on disparate sets of important weights and are processed in distinct parts of the network.

Furthermore, there is an obvious connection between these weight gradient representations and the processes of learning and generalisation. The gradient vector tells us how the objective would change in response to changes in the weights and also the direction in which the weights would be updated if a gradient descent step is taken. Thus, two data points with orthogonal gradient vectors will tend not to generalise to each other: the updates made in training on one point will not affect the other, as they involve distinct sets of weights. Equally, data points with highly aligned gradient vectors will generalise to each other, due to the shared weights.

With these considerations in mind we define our weight-based representations as follows. Given that number agreement is the phenomena we are interested in, we take the objective to be the logit for the singular form of the verb minus the logit for the plural form of the verb. Since our evaluation of models is based on comparing which of these logits is bigger, their difference is a natural, and differentiable, objective to focus on. Furthermore, were we to train on the number agreement task, this or a similar objective would be a reasonable choice.

In particular, if $f(v, p)$ represents the function computed by the network in producing the logit associated with a verb, v, that follows a sentence prefix, p, then the value we are interested, given the sentence *the dog barks*, is the difference, D, between the logits for *barks* and *bark*.

$$D = f(barks, the\ dog) - f(bark, the\ dog) \tag{1}$$

We then take the gradient, $\vec{g}$, of this value with respect to the weights, θ, in both layers of the LSTM, but not the input embeddings or fully connected output layer. These latter parameters are too lexically specific, relating to the particular words in the sentence, while the former make up the general structure processing machinery of the network.

$$\vec{g} = \nabla_\theta D \tag{2}$$

Given two number agreement instances with gradient vectors $\vec{g}_1$ and $\vec{g}_2$, a small gradient directed update based on the first instance produces a change in the weights of $\epsilon\vec{g}_1$ which results in a first order change in singular-plural difference on the second instance of $\epsilon(\vec{g}_1 \cdot \vec{g}_2)$. Thus, the dot product of these vectors is closely related to how learning generalises from one instance to another. However, here we use the cosine as a measure of similarity between the gradients, because it normalises the dot product in terms of the lengths of the vectors, putting all values on a comparable scale, which relates just to the angular alignment of the representations.

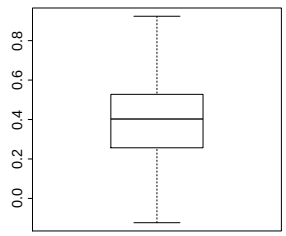
$$sim(\vec{g}_1, \vec{g}_2) = \frac{(\vec{g}_1 \cdot \vec{g}_2)}{\|\vec{g}_1\|\|\vec{g}_2\|} \tag{3}$$

Fort et al. (2020) call such a measure *stiffness*, as it relates to how rigidly changes in the behaviour at one data point are transmitted to another. He and Su (2020) use the term *local elasticity* for a related measure, emphasising instead the flexibility of a model in which the behaviour on a pair of instances is largely independent.

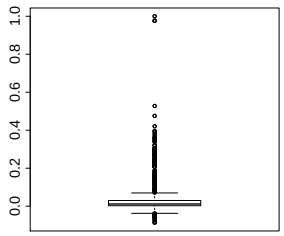
This approach produces a local measure of the which weights are most important in producing a particular number agreement prediction, in the context of small perturbations around the specific values assigned to the weights in the trained model. In contrast, Lakretz et al. (2019) take a more global viewpoint by ablating whole units, and their weights, which can be thought of as comparing performance with the learned weight values to performance when they are set to zero, rather than infinitesimally perturbed. Although this latter approach has the benefit of simplicity, it is a blunter measure and less directly related to learning and generalisation. Nonetheless, the two approaches are related, in the sense that the effect of setting a weight to zero can be expressed in terms of an integral of the gradient along a path that takes the weight to zero.

Here, we use this gradient based similarity measure to compare our model's processing of number agreement across pairs of evaluation sentences. Roughly, positive similarity indicates a pair of sentences that utilise overlapping structures in the model, and across which generalisation is possible. Zero similarity corresponds to orthogonal vectors, indicating a pair without such commonalities, and which are therefore free to behave independently. Negative values are also possible when the shared components end up pushing the pair of outputs in opposite directions, i.e. increasing the output for one data point tends to decrease the output for the other. In our number agreement situation, this would mean something like increasing the probability of a singular verb in one sentence leads to a plural verb becoming more probable in the other.

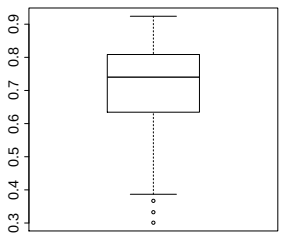
In the next section, we use this approach to compare pairs of number agreement predictions made by the model to understand when these predictions rely on shared architectural resources and when they do



(a) Similarities between different target verbs on the same prefix without reversal.



(b) Similarities of complete unmodified evaluation instances, i.e. different prefix and different target.



(c) Similarities on the same target verb between modified and unmodified versions of the same sentences.

Figure 2: Similarities calculated on the modified model applied to original and modified evaluation sentences. Each data point corresponds to a comparison between weight vectors associated with a pair of number agreement predictions.

not. We begin by examining the original model, comparing both instances that share a common prefix and also instances that are wholly distinct. We then investigate a model trained on modified data, first replicating the results from the original model and finally analysing the relation between natural and unnatural structures.

4.1 Results

We begin by analysing the model trained on the unmodified English corpus in terms of its handling of the sentences in the original evaluation dataset, as a sanity check and baseline against which we compare further results. To this end, we examine the similarity of various singular-plural pairs within the context of the same sentence prefix. That is, if the model has some representation of number features, then it ought to be able, at a minimum, to recognise that the contrast between *barks* and *bark* in the context of *the dog* is similar to the contrast between, for example, *eats* and *eat* in the same context. Thus, we take the full set of pairs of verbs in the evaluation corpus, and use them all in turn to compute singular plural differences in the context of each sentence prefix. We then measure the cosine similarities for distinct verb pairs given the same sentence prefix. That is, we take the cartesian product of all prefixes and target verbs, but measure similarities only between pairs sharing the same prefix.

The boxplot in Figure 1a shows that these similarities are distributed around 0.35 and mainly positive, indicating a reasonable degree of commonality in the weights contributing to the differentiation between singular and plural target verbs, when we compare different targets on the same prefix. This, in turn, suggests generalisation is possible between these different verbs. That is, training on the correct form of a verb following a given prefix will tend to help the model to predict the correct form of a different verb in the context of the same prefix.

This result naturally raises the question of the extent to which such similarities are seen across entirely different sentences. In other words, having shown that the model recognises a connection between *the dog barks* and *the dog eats*, does it also recognise *the dog barks* and *the man eats* as sharing a common structure?

We therefore next compare gradient vectors for whole evaluation instances, measuring the extent to which distinct targets - e.g. *barks* vs. *eats* - are aligned in the context of distinct prefixes - e.g. *the dog* vs. *the man*. Thus, the similarities are measured between number agreement predictions for the same set of verbs as the previous analysis, but only with their original prefixes from the evaluation dataset. The resulting similarities in Figure 1b are distributed mainly around zero, suggesting that without a common prefix, the processing of two evaluation instances relies on disparate sets of weights. In other words, the network lacks a generic mechanism for handling number agreement, instead relying on different weights to handle different cases. If this is accurate, then large numbers of training sentences are needed to cover these various cases and estimate the relevant weights. This would also mean that generalisation to the

evaluation instances happens, not in in terms of a general number agreement mechanism, but in terms of being sufficiently similar to a training instance.

The datapoints plotted in Figure 1b include comparisons which agree in number - e.g. both singular: *the man eats* and *the dog barks* - alongside comparisons which disagree - e.g. singular vs. plural: *the man eats* and *the cats meow*. If we break this distribution of similarities down by the correct number in each instance pair, i.e. whether we are comparing a singular prefix to a plural prefix, etc., then, as seen in Figure 1c, we find that all these comparisons produce roughly similar distributions of cosines.

The discordant comparisons, SP and PS, in which we compare a singular to a plural instance and vice versa, do produce more negative values than the concordant comparisons, SS and PP, but the proportion of positive values is greater than 75% in all four cases. These positive values mean that training on a singular verb form makes singular verb forms more likely across the board, including those cases where plural is the correct form. If training on a correct singular form generalised to more correct plural forms, then the SP similarities would be negative. Thus, these positive similarities indicate competition, rather than generalisation, between singular and plural instances. In other words, the network appears to lack an abstract concept of number agreement that connects both singular and plural agreement into a unified whole.

We now turn to analysing a model trained on modified data, focusing on the reversed order model, as its performance was closest to the original and it is the most linguistically interesting transformation. Figure 2a reproduces the same baseline analysis as before on the new model using the unmodified evaluation corpus. In other words, we measure cosine similarities between different verbs on the same prefix, without reversing any part of the sentence. The boxplot confirms that the modified model performs in a comparable manner to the original model. Similarly, the similarities in Figure 2b from comparing whole evaluation instances are distributed mainly around a value close to zero, in the same manner as those for the original model.

In this case, we can also compare original sentences to their modified form, to investigate whether the reversed and non-reversed elements of the language share common resources in the network. For example, we would compare the number agreement prediction on *bark* vs *barks* across the inputs *The man whose dog barks likes apples* and *The man whose dog* $\langle marker \rangle$ *apples likes barks*. The boxplot in Figure 2c shows that these similarities are distributed around 0.7, making them more similar even than the baseline similarities. On one hand, these high similarities are consistent with the fact these comparisons are made on the same target, whereas the baseline compares different targets. On the other hand, the target is now embedded in a radically different structure in which word orders are reversed and dependencies are tangled.

The experiments of Musso et al. (2003) and Smith et al. (1993) suggest that when humans attempt to learn such unnatural structures, they have to employ distinct cognitive resources and brain regions, whereas the network appears to be sharing substantial common weights between the modified and unmodified forms. In fact, according to the distributions of similarities in Figures 2a and Figure 2c, the natural structure has more in common with the unnatural structure than it does with other natural structures. Thus, the networks encoding of knowledge into its weights is quite unlike the acquisition of natural languages by humans.

4.2 Discussion

These results are consistent with those of Lakretz et al. (2019), who found that long and short range dependencies were handled by distinct sets of units and that the influence of the long range units varied substantially between different types of constructions. In other words, both sets of results indicate that number agreement phenomena are processed in multiple, disparate parts of the network, depending on the particular context they occur in. Thus, rather than treating these grammatical structures in a unified manner, the network handles them as a collection of separate phenomena which must be learned independently. In terms of the unnatural languages, the ability to segregate the data into local non-interacting neighbourhoods supports the flexibility of the model in handling the multiple contextually dependent structures, such as original and reversed word orders, in a single model.

The excess capacity discussed in Section 3.2 is also reflected here in this distribution of grammatical knowledge across a number of disparate places in the model, rather than in a smaller number of shared weights. In fact, recent research (Belkin et al., 2019; Nakkiran et al., 2020) suggests that pushing a network's capacity beyond the point at which it can memorise the entire dataset may help the model to interpolate more smoothly between training examples. However, this is unlikely to produce human-like learning and generalisation, particularly in terms of extrapolation beyond the training distribution. For example, Kim and Linzen (2020) find that neural models of semantic parsing struggle to generalise from shallower - e.g. *Ava saw the ball in the bottle on the table* - to more deeply nested structures - e.g. *Ava saw the ball in the bottle on the table on the floor* - in a consistent and unified manner.

In our case, learning an effective and unified concept of number agreement, embodied in a single shared architectural component, probably requires the right sort of inductive biases, attuned to the sort of hierarchical dependency structures that such agreement happens within. As we have seen, the LSTM is equally happy to learn normal linguistic dependencies, pathological dependencies or even no dependencies. Moreover, the observed similarities for natural and unnatural structures is an example of sharing resources when they should be differentiated. A more psycholinguistically plausible architecture may need to be innately sensitive to these differences.

5 Conclusions

The LSTM architecture of Gulordava et al. (2018) demonstrated its capacity in our experiments to handle structures that go beyond those seen in natural languages. Seen as a test of broad coverage sequence learning, these results are evidence of its flexibility and generality. However, seen from the perspective of modelling the shared design principles of human languages, these characteristics begin to look like weaknesses. They suggest that the network lacks the necessary inductive biases to distinguish between natural and unnatural structures.

Many linguists interpret these shared principles in terms of innate cognitive abilities (Chomsky, 1986; Bickerton, 1984), with additional evidence coming from human experiments with artificial languages, such as those described in Section 2. Furthermore, the behaviour of the neural models on out-of-domain examples also suggests stronger innate biases or constraints may be useful.

Although neural NLP systems often achieve impressive in-domain performance, there is still a problem with the robustness of these abilities (Jia and Liang, 2017; Belinkov and Bisk, 2018; Carmona et al., 2018). Making the internal mechanisms of our architectures more human-like is a plausible direction from which to address these problems. Moreover, improving the specificity of these models for natural languages, as opposed to arbitrary sequential structures, should reduce the problem of requiring huge datasets to train huge numbers of weights.

Our analysis of how knowledge is encoded in the weights of the model also underscores the non-human-like form of learning embodied in this architecture. The network appears to lack a shared set of weights that could instantiate a unifying concept of number agreement to tie together the disparate predictions of singular versus plural it makes for particular constructions. Similarly, Lakretz et al. (2019) found that long and short range dependencies were handled by independent units. At the same time, the model also appears to process natural and unnatural structures using common resources.

Future work will examine how to remedy these shortcomings, by examining how inductive biases change the way in which information is encoded in the weights during learning. For example, Mitchell and Bowers (2020) explore how a stack-like structure can be imposed on the memory cells of an LSTM that shares weights between long and short range dependencies, obtaining improved generalisation on a context free language learning task.

Acknowledgements

This research was supported by the European Research Council Grant Generalization in Mind and Machine, ID number 741134. The authors would like to thank Nina Kazanina, Conor Houghton and the Memory and Language group in Bristol for discussion and feedback.

References

Aristotle. 350 BCE. *On Interpretation*. Medieval Philosophical Texts in Translation. Marquette University Press.

Peter L. Bartlett and Shahar Mendelson. 2003. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463482, March.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Derek Bickerton. 1984. The language bioprogram hypothesis. *Behavioral and Brain Sciences*, 7(2):173188.

Daniel M. Bikel. 2004. Intricacies of Collins' parsing model. *Computational Linguistics*, 30(4):479–511.

Vicente Iván Sánchez Carmona, Jeff Mitchell, and Sebastian Riedel. 2018. Behavior analysis of nli models: Uncovering the influence of three factors on robustness. In *NAACL-HLT*, pages 1975–1985.

Noam Chomsky. 1956. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton and Co., The Hague.

Noam Chomsky. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Convergence (New York, N.Y.). Praeger.

Michael John Collins. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL 96, page 184191, USA. Association for Computational Linguistics.

Marin Dujmović, Gaurav Malhotra, and Jeffrey S Bowers. 2020. What do adversarial images tell us about human vision? *eLife*, 9:e55978, September.

Sandiway Fong and Robert C. Berwick. 2008. Treebank parsing and knowledge of language: A cognitive perspective. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*.

Sandiway Fong, Igor Malioutov, Beracah Yankama, and Robert C. Berwick. 2013. Treebank parsing and knowledge of language. In Aline Villavicencio, Thierry Poibeau, Anna Korhonen, and Afra Alishahi, editors, *Cognitive Aspects of Computational Language Acquisition*, Theory and Applications of Natural Language Processing, pages 133–172. Springer.

Stanislav Fort, Paweł Krzysztof Nowak, Stanisław Jastrzebski, and Srini Narayanan. 2020. Stiffness: A new perspective on generalization in neural networks.

Stefan L. Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834.

Stefan L. Frank, Rens Bod, and Morten H. Christiansen. 2012. How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, 279(1747):4522–4531.

Stefan L. Frank, L.J. Otten, G. Galli, and G. Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1 – 11.

Stefan L. Frank. 2009. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hangfeng He and Weijie Su. 2020. The local elasticity of neural networks. In *International Conference on Learning Representations*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):17351780, November.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, September. Association for Computational Linguistics.

A. K. Joshi, S. R. Kosaraju, and H. Yamada. 1969. String adjunct grammars. In *10th Annual Symposium on Switching and Automata Theory (swat 1969)*, pages 245–262.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July. Association for Computational Linguistics.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 11–20, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jeff Mitchell and Jeffrey S. Bowers. 2020. Harnessing the symmetry of convolutions for systematic generalisation. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–8. IEEE.

Mariacristina Musso, Andrea Moro, Volkmar Glauche, Michel Rijntjes, Juergen Reichenbach, Christian Büchel, and Cornelius Weiller. 2003. Broca's area and the language instinct. *Nature Neuroscience*, 6:774–781.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2020. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Pāṇini. 6th Century BCE. *Aṣṭādhyāyī of Pāṇini*. Texas linguistics series. Motilal Banarsidass.

Joe Pater. 2019. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, 95(1):e41–e74.

Steven Pinker and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1):73 – 193.

Jonathan Rawski and Jeffrey Heinz. 2019. No free lunch in linguistics or machine learning: Response to Pater. *Language*, 95(1).

D. E. Rumelhart and J. L. McClelland, 1986. *On Learning the Past Tenses of English Verbs*, page 216271. MIT Press, Cambridge, MA, USA.

Neil V. Smith, Ianthi-Maria Tsimpli, and Jamal Ouhalla. 1993. Learning the impossible: The acquisition of possible and impossible languages by a polyglot savant. *Lingua*, 91(4):279 – 347.

Mark Steedman. 1987. Combinatory grammars and parasitic gaps. *Natural Language & Linguistic Theory*, 5:403–439.

D. H. Wolpert and W. G. Macready. 1997. No free lunch theorems for optimization. *Trans. Evol. Comp*, 1(1):6782, April.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.