Colorless green recurrent networks dream hierarchically

Kristina Gulordava[*]
Department of Linguistics
University of Geneva
kristina.gulordava@unige.ch

Piotr Bojanowski
Facebook AI Research
Paris
bojanowski@fb.com

Edouard Grave
Facebook AI Research
New York
egrave@fb.com

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

Marco Baroni
Facebook AI Research
Paris
mbaroni@fb.com

Abstract

Recurrent neural networks (RNNs) have achieved impressive results in a variety of linguistic processing tasks, suggesting that they can induce non-trivial properties of language. We investigate here to what extent RNNs learn to track abstract hierarchical syntactic structure. We test whether RNNs trained with a generic language modeling objective in four languages (Italian, English, Hebrew, Russian) can predict long-distance number agreement in various constructions. We include in our evaluation nonsensical sentences where RNNs cannot rely on semantic or lexical cues ("The colorless green ideas I ate with the chair sleep furiously"), and, for Italian, we compare model performance to human intuitions. Our language-model-trained RNNs make reliable predictions about long-distance agreement, and do not lag much behind human performance. We thus bring support to the hypothesis that RNNs are not just shallow-pattern extractors, but they also acquire deeper grammatical competence.

1 Introduction

Recurrent neural networks (RNNs; Elman, 1990) are general sequence processing devices that do not explicitly encode the hierarchical structure that is thought to be essential to natural language (Everaert et al., 2015). Early work using artificial languages showed that they may nevertheless be able to approximate context-free languages (Elman, 1991). More recently, RNNs have achieved impressive results in large-scale tasks such as language modeling for speech recognition and machine translation, and are by now standard tools for sequential natural language tasks (e.g., Mikolov et al., 2010; Graves, 2012; Wu et al., 2016). This suggests that RNNs may learn to track grammatical structure even when trained on noisier natural data. The conjecture is supported by the success of RNNs as feature extractors for syntactic parsing (e.g., Cross and Huang, 2016; Kiperwasser and Goldberg, 2016; Zhang et al., 2017).

Linzen et al. (2016) directly evaluated the extent to which RNNs can approximate hierarchical structure in corpus-extracted natural language data. They tested whether RNNs can learn to predict English subject-verb agreement, a task thought to require hierarchical structure in the general case ("the girl the boys like... is or are?"). Their experiments confirmed that RNNs can, in principle, handle such constructions. However, in their study RNNs could only succeed when provided with explicit supervision on the target task. Linzen and colleagues argued that the unsupervised language modeling objective is not sufficient for RNNs to induce the syntactic knowledge necessary to cope with long-distance agreement.

The current paper reevaluates these conclusions. We strengthen the evaluation paradigm of Linzen and colleagues in several ways. Most importantly, their analysis did not rule out the possibility that RNNs might be relying on *semantic* or *collocational/frequency-based* information, rather than purely on syntactic structure. In "dogs in the neighbourhood often bark", an RNN might get the right agreement by encoding information

[*]The work was conducted during the internship at Facebook AI Research, Paris.

Proceedings of NAACL-HLT 2018, pages 1195–1205
New Orleans, Louisiana, June 1 - 6, 2018.

about what typically barks (dogs, not neighbourhoods), without relying on more abstract structural cues. In a follow-up study to Linzen and colleagues', Bernardy and Lappin (2017) observed that RNNs are better at long-distance agreement when they construct rich lexical representations of words, which suggests effects of this sort might indeed be at play.

We introduce a method to probe the syntactic abilities of RNNs that abstracts away from potential lexical, semantic and frequency-based confounds. Inspired by Chomsky's (1957) insight that "grammaticalness cannot be identified with meaningfulness" (p. 106), we test long-distance agreement both in standard corpus-extracted examples and in comparable nonce sentences that are grammatical but completely meaningless, e.g., (paraphrasing Chomsky): "The colorless green ideas I ate with the chair sleep furiously".

We extend the previous work in three additional ways. First, alongside English, which has few morphological cues to agreement, we examine Italian, Hebrew and Russian, which have richer morphological systems. Second, we go beyond subject-verb agreement and develop an automated method to harvest a variety of long-distance number agreement constructions from treebanks. Finally, for Italian, we collect human judgments for the tested sentences, providing an important comparison point for RNN performance.[1]

We focus on the more interesting unsupervised setup, where RNNs are trained to perform generic, large-scale language modeling (LM): they are not given explicit evidence, at training time, that they must focus on long-distance agreement, but they are rather required to track a multitude of cues that might help with word prediction in general.

Our results are encouraging. RNNs trained with a LM objective solve the long-distance agreement problem well, even on nonce sentences. The pattern is consistent across languages, and, crucially, not far from human performance in Italian. Moreover, RNN performance on language modeling (measured in terms of perplexity) is a good predictor of long-distance agreement accuracy. This suggests that the ability to capture structural generalizations is an important aspect of what makes the best RNN architectures so good at language modeling. Since our positive results contradict, to some extent, those of Linzen et al. (2016), we also replicate their relevant experiment using our best RNN (an LSTM). We outperform their models, suggesting that a careful architecture/hyperparameter search is crucial to obtain RNNs that are not only good at language modeling, but able to extract syntactic generalizations.

[1]The code to reproduce our experiments and the data used for training and evaluation, including the human judgments in Italian, can be found at https://github.com/facebookresearch/colorlessgreenRNNs.

2 Constructing a long-distance agreement benchmark

Overview. We construct our number agreement test sets as follows. **Original** sentences are automatically extracted from a dependency treebank. They are then converted into **nonce** sentences by substituting all content words with random words with the same morphology, resulting in grammatical but nonsensical sequences. An LM is evaluated on its predictions for the target (second) word in the dependency, in both the original and nonce sentences.

Long-distance agreement constructions. Agreement relations, such as subject-verb agreement in English, are an ideal test bed for the syntactic abilities of LMs, because the form of the second item (the target) is predictable from the first item (the cue). Crucially, the cue and the target are linked by a *structural* relation, where linear order in the word sequence does not matter (Everaert et al., 2015). Consider the following subject-verb agreement examples: "the girl thinks…", "the girl [you met] thinks…", "the girl [you met yesterday] thinks…", "the girl [you met yesterday through her friends] thinks…". In all these cases, the number of the main verb "thinks" is determined by its subject ("girl"), and this relation depends on the syntactic structure of the sentence, not on the linear sequence of words. As the last sentence shows, the word directly preceding the verb can even be a noun with the opposite number ("friends"), but this does not influence the structurally-determined form of the verb.

When the cue and the target are adjacent ("the girl thinks…"), an LM can predict the target without access to syntactic structure: it can simply extract the relevant morphosyntactic features of words (e.g., number) and record the co-occurrence frequencies of patterns such as N_{Plur} V_{Plur} (Mikolov et al., 2013). Thus, we focus here on *long-distance* agreement, where an arbitrary num-

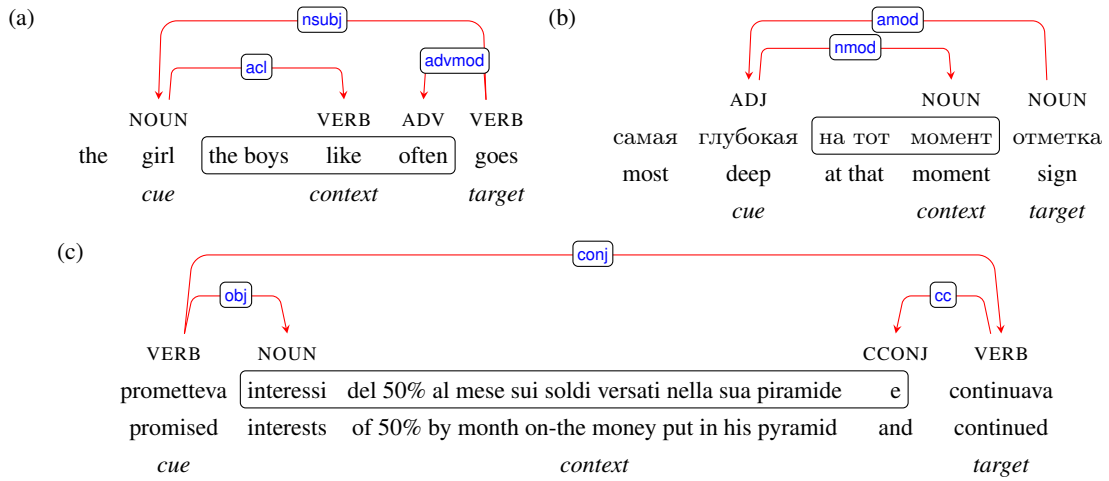

Figure 1: Example agreement constructions defined by a dependency and the separating context, in (a) English, (b) Russian and (c) Italian.

ber of words can occur between the elements of the agreement relation. We limit ourselves to *number* agreement (plural or singular), as it is the only overt agreement feature shared by all of the languages we study.

Identifying candidate constructions. We started by collecting pairs of part-of-speech (POS) tags connected by a dependency arc. Independently of which element is the head of the relation, we refer to the first item as the **cue** and to the second as the **target**. We additionally refer to the POS sequence characterizing the entire pattern as a **construction**, and to the elements in the middle as **context**.

For each candidate construction, we collected all of the contexts in the corpus that intervene between the cue and the target (we define contexts as the sequence of POS tags of the top-level nodes in the dependency subtrees). For example, for the English subject-verb agreement construction shown in Fig. 1a, the context is defined by VERB (head of the relative clause) and ADV (adverbial modifier of the target verb), which together dominate the sequence "the boys like often". For the Russian adjective-noun agreement construction in Fig. 1b, the context is NOUN, because in the dependency grammar we use the noun "moment" is the head of the prepositional phrase "at that moment", which modifies the adjective "deep". The candidate agreement pair and the context form a construction, which is characterized by a sequence of POS tags, e.g., NOUN VERB ADV VERB or VERB NOUN CCONJ VERB (Fig. 1c).

Our constructions do not necessarily correspond to standard syntactic structures. The English subject-verb agreement construction NOUN VERB VERB, for example, matches both object and subject relative clause contexts, e.g., "girl the boys like is" and "girls who stayed at home were". Conversely, standard syntactic structures might be split between different constructions, e.g., relative clause contexts occur in both NOUN VERB VERB and NOUN VERB ADV VERB constructions (the latter is illustrated by the English example in Fig. 1a).

Construction contexts can contain a variable numbers of words. Since we are interested in challenging cases, we only considered cases in which at least three tokens intervened between the cue and the target.

Excluding non-agreement constructions. In the next step, we excluded constructions in which the candidate cue and target did not agree in number in all of the instances of the construction in the treebank (if both the cue and the target were morphologically annotated for number). This step retained English subject-verb constructions, for example, but excluded verb-object constructions, since any form of a verb can appear both with singular and plural objects. To focus on robust agreement patterns, we only kept constructions with at least 10 instances of both plural and singular agreement.

When applied to the treebanks we used (see Section 3), this step resulted in between two (English) and 21 (Russian) constructions per lan-

guage. English has the poorest morphology and consequently the lowest number of patterns with identifiable morphological agreement. Only the VP-conjunction construction (Fig. 1c) was identified in all four languages. Subject-verb agreement constructions were extracted in all languages but Russian; Russian has relatively flexible word order and a noun dependent preceding a head verb is not necessarily its subject. The full list of extracted constructions in English and Italian is given in Tables 2 and 3, respectively. For the other languages, see the Supplementary Material (SM).[2]

Original sentence test set. Our "original" sentence test set included all sentences from each construction where all words from the cue and up to and including the target occurred in the LM vocabulary (Section 3), and where the singular/plural counterpart of the target occurred in the treebank and in the language model vocabulary (this is required by the evaluation procedure outlined below). The total counts of constructions and original sentences in our test sets are provided in Table 1. The average number of context words separating the cue and the target ranged from 3.6 (Hebrew) to 4.5 (Italian).

Generating nonce sentences. We generated nine nonce variants of each original sentence as follows. Each content word (noun, verb, adjective, proper noun, numeral, adverb) in the sentence was substituted by another random content word from the treebank with matching POS and morphological features. To avoid forms that are ambiguous between several POS, which are particularly frequent in English (e.g., plural noun and singular verb forms), we excluded the forms that appeared with a different POS more than 10% of the time in the treebank. Function words (determiners, pronouns, adpositions, particles) and punctuation were left intact. For example, we generated the nonce (1b) from the original sentence (1a):

(1) a. It presents the case for marriage equality and states...
b. It stays the shuttle for honesty insurance and finds...

Note that our generation procedure is based on morphological features and does not guarantee that argument structure constraints are respected (e.g., "it stays the shuttle" in (1b)).

Evaluation procedure. For each sentence in our test set, we retrieved from our treebank the form that is identical to the agreement target in all morphological features except number (e.g., "finds" instead of "find" in (1b)). Given a sentence with prefix p up to and excluding the target, we then compute the probabilities $P(t_1|p)$ and $P(t_2|p)$ for the singular and plural variants of the target, t_1 and t_2, based on the language model. Following Linzen et al. (2016), we say that the model identified the correct target if it assigned a higher probability to the form with the correct number. In (1b), for example, the model should assign a higher probability to "finds" than "find".[3]

3 Experimental setup

Treebanks. We extracted our test sets from the Italian, English, Hebrew and Russian Universal Dependency treebanks (UD, v2.0, Nivre et al., 2016). The English and Hebrew treebanks were post-processed to obtain a richer morphological annotation at the word level (see SM for details).

LM training data. Training data for Italian, English and Russian were extracted from the respective Wikipedias. We downloaded recent dumps, extracted the raw text from them using WikiExtractor[4] and tokenized it with TreeTagger (Schmid, 1995). We also used the TreeTagger lemma annotation to filter out sentences with more than 5% unknown words. For Hebrew, we used the preprocessed Wikipedia corpus made available by Yoav Goldberg.[5] We extracted 90M token subsets for each language, shuffled them by sentence and split them into training and validation sets (8-to-1 proportion). For LM training, we included the 50K most frequent words in each corpus in the vocabulary, replacing the other tokens with the UNK symbol. The validation set perplexity values we report below exclude unknown tokens.

RNN language models. We experimented with simple RNNs (sRNNs, Elman, 1990), and their most successful variant, long-short term memory models (LSTMs, Hochreiter and Schmidhu-

[2] The SM is available as a standalone file on the project's public repository.

[3] Obviously, in the nonce cases, the LMs never assigned the highest overall probability to either of the two candidates. Qualitatively, in such cases LMs assigned the largest absolute probabilities to plausible frequent words.

[4] https://github.com/attardi/wikiextractor

[5] http://u.cs.biu.ac.il/~yogo/hebwiki/

ber, 1997). We use the PyTorch RNN implementation.[6] We trained the models with two hidden layer dimensionalities (650 and 200 units), and a range of batch sizes, learning rates and dropout rates. See SM for details on hyperparameter tuning. In general, a larger hidden layer size was the best predictor of lower perplexity. Given that our LSTMs outperformed our sRNNs, our discussion of the results will focus on the former; we will use the terms LSTM and RNN interchangeably.[7]

Baselines. We consider three baselines: first, a **unigram** baseline, which picks the most frequent form in the training corpus out of the two candidate target forms (singular or plural); second, a 5-gram model with Kneser-Ney smoothing (**KN**, Kneser and Ney, 1995) trained using the IRSTLM package (Federico et al., 2008) and queried using KenLM (Heafield, 2011); and third, a **5-gram LSTM**, which only had access to windows of five tokens (Chelba et al., 2017). Compared to KN, the 5-gram LSTM can generalize to unseen n-grams thanks to its embedding layer and recurrent connections. However, it cannot discover long-distance dependency patterns that span more than five words. See SM for details on the hyperparameters of this baseline.

Human experiment in Italian. We presented the full Italian test set (119 original and 1071 nonce sentences) to human subjects through the Amazon Mechanical Turk interface.[8] We picked Italian because, being morphologically richer, it features more varied long-distance constructions than English. Subjects were requested to be native Italian speakers. They were presented with a sentence up to and excluding the target. The singular and plural forms of the target were presented below the sentence (in random order), and subjects were asked to select the more plausible form.

To prevent long-distance agreement patterns from being too salient, we mixed the test set with the same number of filler sentences. We started from original fillers, which were random treebank-extracted sentences up to a content word in singular or plural form. We then generated nonce fillers from the original ones using the procedure outlined in Section 2. A control subset of 688 fillers was manually selected by a linguistically-trained Italian native speaker as unambiguous cases. To make sure we were only using data from native (or at least highly proficient) Italian speakers, we filtered out the responses of subjects who chose the wrong target in more than 20% of the fillers.

We collected on average 9.5 judgments for each item (minimum 5 judgments). To account for the variable number of judgments across sentences, accuracy rates were first calculated within each sentence and then averaged across sentences.

[6] https://github.com/pytorch/examples/tree/master/word_language_model

[7] Detailed results for sRNNs can be found in the SM.

[8] https://www.mturk.com/

	IT	EN	HE	RU
#constructions	8	2	18	21
#original	119	41	373	442
Unigram				
Original	54.6	65.9	67.8	60.2
Nonce	54.1	42.5	63.1	54.0
5-gram KN				
Original	63.9	63.4	72.1	73.5
Nonce	52.8	43.4	61.7	56.8
Perplexity	147.8	168.9	122.0	166.6
5-gram LSTM				
Original	81.8 ±3.2	70.2 ±5.8	90.9 ±1.2	91.5 ±0.4
Nonce	78.0 ±1.3	58.2 ±2.1	77.5 ±0.8	85.7 ±0.7
Perplexity	62.6 ±0.2	71.6 ±0.3	59.9 ±0.2	61.1 ±0.4
LSTM				
Original	92.1 ±1.6	81.0 ±2.0	94.7 ±0.4	96.1 ±0.7
Nonce	85.5 ±0.7	74.1 ±1.6	80.8 ±0.8	88.8 ±0.9
Perplexity	45.2 ±0.3	52.1 ±0.3	42.5 ±0.2	48.9 ±0.6

Table 1: Experimental results for all languages averaged across the five best models in terms of perplexity on the validation set. Original/Nonce rows report percentage accuracy, and the numbers in small print represent standard deviation within the five best models.

4 Results

The overall results are reported in Table 1. We report results averaged across the five models with the lowest validation perplexity, as well as standard deviations across these models. In summary,

		N V V	V NP conj V
Italian	Original	$93.3_{\pm 4.1}$	$83.3_{\pm 10.4}$
	Nonce	$92.5_{\pm 2.1}$	$78.5_{\pm 1.7}$
English	Original	$89.6_{\pm 3.6}$	$67.5_{\pm 5.2}$
	Nonce	$68.7_{\pm 0.9}$	$82.5_{\pm 4.8}$
Hebrew	Original	$86.7_{\pm 9.3}$	$83.3_{\pm 5.9}$
	Nonce	$65.7_{\pm 4.1}$	$83.1_{\pm 2.8}$
Russian	Original	-	$95.2_{\pm 1.9}$
	Nonce	-	$86.7_{\pm 1.6}$

Table 2: LSTM accuracy in the constructions N V V (subject-verb agreement with an intervening embedded clause) and V NP conj V (agreement between conjoined verbs separated by a complement of the first verb).

the LSTM clearly outperformed the other LMs. Rather surprisingly, its performance on nonce sentences was only moderately lower than on original ones; in Italian this gap was only 6.6%.

The KN LM performed poorly; its accuracy on nonce sentences was comparable to that of the unigram baseline. This confirms that the number of the target in nonce sentences cannot be captured by shallow n-gram patterns. The 5-gram LSTM model greatly improved over the KN baseline; its accuracy dropped only modestly between the original and nonce sentences, demonstrating its syntactic generalization ability. Still, the results are substantially below those of the LSTM with unlimited history. This confirms that our test set contains hard long-distance agreement dependencies, and, more importantly, that the more general LSTM model can exploit broader contexts to learn about and track long-distance syntactic relations.

The increase in accuracy scores across the three LMs (KN, 5-gram LSTM and unbounded-context LSTM) correlates well with their validation perplexities in the language modeling task. We also found a strong correlation between agreement accuracy and validation perplexity across all the LSTM variants we explored in the hyperparameter search (68 models per language), with Pearson correlation coefficients ranging from $r = -0.55$ in Hebrew to $r = -0.78$ in English ($p < 0.001$ in all languages). This suggests that acquiring abstract syntactic competence is a natural component of the skills that improve the generic language modeling performance of RNNs.

Differences across languages. English was by far the hardest language. We conjecture that this is due to its poorer morphology and higher POS ambiguity, which might not encourage a generic language model to track abstract syntactic configurations. There is an alternative hypothesis, however. We only extracted two constructions for English, both of which can be argued to be linguistically complex: subject-verb agreement with an intervening embedded clause, and agreement between two conjoined verbs with a nominal complement intervening between the verbs. Yet the results on these two constructions, comparable across languages (with the exception of the subject-verb construction in Russian, which was not extracted), confirm that English is particularly hard (Table 2). A qualitative inspection suggests that the low accuracy in the verb conjunction case (67.5%) is due to ambiguous sentences such as "if you have any questions or need/needs", where the target can be re-interpreted as a noun that is acceptable in the relevant context.[9]

In languages such as Italian and Russian, which have richer morphology and less ambiguity at the part-of-speech level than English, the LSTMs show much better accuracy and a smaller gap between original and nonce sentences. These results are in line with human experimental studies that found that richer morphology correlates with fewer agreement attraction errors (Lorimor et al., 2008). The pattern of accuracy rates in general, and the accuracy for the shared V NP conj V construction in particular, are consistent with the finding that Russian is less prone to human attraction errors than Italian, which, in turn, shows less errors than English.

The largest drop in accuracy between original and nonce sentences occurred in Hebrew. A qualitative analysis of the data in this language suggests that this might be due to the numerical prevalence of a few constructions that can have multiple alternative readings, some of which can license the incorrect number. We leave a more systematic analysis of this finding for future research.

Human results. To put our results in context and provide a reasonable upper bound on the LM performance, in particular for nonce sentences, we next compare model performance to that of human

[9]The nonce condition has higher accuracy because our substitution procedure in English tends to reduce POS ambiguity.

Construction	#original	Original		Nonce	
		Subjects	LSTM	Subjects	LSTM
DET [AdjP] NOUN	14	98.7	$98.6_{\pm 3.2}$	98.1	$91.7_{\pm 0.4}$
NOUN [RelC / PartP] clitic VERB	6	93.1	$100_{\pm 0.0}$	95.4	$97.8_{\pm 0.8}$
NOUN [RelC / PartP] VERB	27	97.0	$93.3_{\pm 4.1}$	92.3	$92.5_{\pm 2.1}$
ADJ [conjoined ADJs] ADJ	13	98.5	$100_{\pm 0.0}$	98.0	$98.1_{\pm 1.1}$
NOUN [AdjP] relpron VERB	10	95.9	$98.0_{\pm 4.5}$	89.5	$84.0_{\pm 3.3}$
NOUN [PP] ADVERB ADJ	13	91.5	$98.5_{\pm 3.4}$	79.4	$76.9_{\pm 1.4}$
NOUN [PP] VERB (participial)	18	87.1	$77.8_{\pm 3.9}$	73.4	$71.1_{\pm 3.3}$
VERB [NP] CONJ VERB	18	94.0	$83.3_{\pm 10.4}$	86.8	$78.5_{\pm 1.7}$
(Micro) average		94.5	$92.1_{\pm 1.6}$	88.4	$85.5_{\pm 0.7}$

Table 3: Subject and LSTM accuracy on the Italian test set, by construction and averaged.

subjects in Italian.

Table 3 reports the accuracy of the LSTMs and the human subjects, grouped by construction.[10] There was a consistent gap in human accuracy between original and nonce sentences (6.1% on average). The gap in accuracy between the human subjects and the model was quite small, and was similar for original and nonce sentences (2.4% and 2.9%, respectively).

In some of the harder constructions, particularly subject-verb agreement with an embedded clause, the accuracy of the LSTMs on nonce sentences was comparable to human accuracy ($92.5_{\pm 2.1}$ vs. 92.3%). To test whether the human subjects and the models struggle with the same sentences, we computed for each sentence (1) the number of times the human subjects selected the correct form of the target minus the number of times they selected the incorrect form, and (2) the difference in model log probability between the correct and incorrect form. The Spearman correlation between these quantities was significant, for both original ($p < 0.05$) and nonce sentences ($p < 0.001$). This indicates that humans were more likely to select the correct form in sentences in which the models were more confident in a correct prediction.

Moreover, some of the easiest and hardest constructions are the same for the human subjects and the models. In the easy constructions DET [AdjP] NOUN[11] and ADJ [conjoined ADJs] ADJ, one or more adjectives that intervene between the cue and the target agree in number with the target, providing shorter-distance evidence about its correct number. For example, in

(2) un film inutile ma almeno festivo e
a movie useless but at.least festive and
giovanile
youthful
"A useless but at least festive and youthful movie"

the adjective "festivo" is marked for singular number, offering a nearer reference for the target number than the cue "inutile". At the other end, NOUN [PP] VERB (participial) and NOUN [PP] ADVERB ADJ are difficult. Particularly in the nonce condition, where semantics is unhelpful or even misleading, the target could easily be interpreted as a modifier of the noun embedded in the preceding prepositional phrase. For example, for the nonce case:

(3) orto di regolamenti davvero pedonale/i
orchard of rules truly pedestrian
"truly pedestrian orchard of rules"

both the subjects and the model preferred to treat "pedestrian" as a modifier of "rules" ("orchard of truly pedestrian rules"), resulting in the wrong agreement given the intended syntactic structure.

Attractors. We define attractors as words with the same POS as the cue but the opposite number, which intervene in the linear order of the sen-

[10] The SM contains the results for the other languages broken down by construction. Note that Table 3 reports linguistically intuitive construction labels. The corresponding POS patterns are (in same order as table rows): DET ADJ NOUN, NOUN VERB PRON VERB, NOUN VERB VERB, ADJ ADJ CCONJ ADJ, NOUN ADJ PUNCT PRON VERB, NOUN NOUN ADV ADJ, NOUN NOUN VERB, VERB NOUN CCONJ VERB.

[11] The relatively low nonce LSTM performance on this construction is due to a few adjectives that could be reinterpreted as nouns.

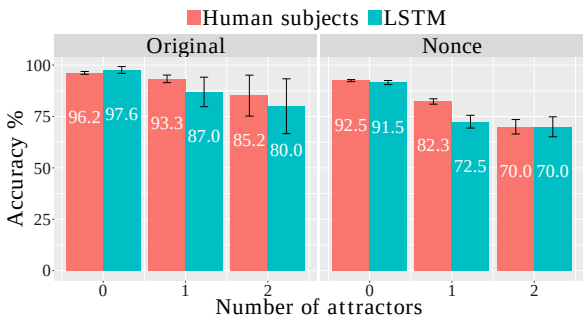

Figure 2: Accuracy by number of attractors in Italian. Human performance is shown in red and LSTM in blue (median model among top 5 ranked by perplexity). Error bars show standard error.

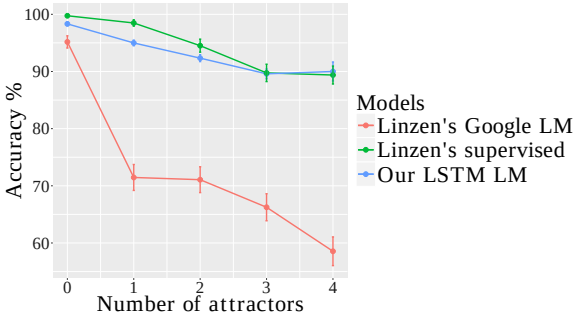

Figure 3: Linzen's attractor set. Our LM-trained LSTM (blue; "median" model) compared to their LSTM with explicit number supervision (green) and their best LM-trained LSTM (red).

tence between the cue and the target. Attractors constitute an obvious challenge for agreement processing (Bock and Miller, 1991). We show how their presence affects human and model behavior in Fig. 2. We limit our analysis to a maximum of two attractors, since there were only two original sentences in the test corpus with three attractors or more. Both model and human accuracies degraded with the number of attractors; the drop in accuracy was sharper in the nonce condition. While the model performed somewhat worse than humans, the overall pattern was comparable.

Our results suggest that the LSTM is quite robust to the presence of attractors, in contrast to what was reported by Linzen et al. (2016). We directly compared our English LSTM LM to theirs by predicting verb number on the Linzen et al. (2016) test set. We extracted sentences where all of the words between subject and verb were in our LM vocabulary. Out of those sentences, we sampled 2000 sentences with 0, 1 and 2 attractors and kept all the sentences with 3 and 4 attractors (1329 and 347 sentences, respectively). To ensure that our training set and Linzen's test set do not overlap (both are based on Wikipedia texts), we filtered out all of test sentences that appeared in our training data (187 sentences).

Fig. 3 compares our results to the results of the best LM-trained model in Linzen et al. (2016) (their "Google LM").[12] Not only did our LM greatly outperform theirs, but it approached the performance of their supervised model.[13] This difference in results points to the importance of careful tuning of LM-trained LSTMs, although we must leave to a further study a more detailed understanding of which differences crucially determine our better performance.

5 Related work

Early work showed that RNNs can, to a certain degree, handle data generated by context-free and even context-sensitive grammars (e.g., Elman, 1991, 1993; Rohde and Plaut, 1997; Christiansen and Chater, 1999; Gers and Schmidhuber, 2001; Cartling, 2008). These experiments were based on small and controlled artificial languages, in which complex hierarchical phenomena were often overrepresented compared to natural languages.

Our work, which is based on naturally occurring data, is most closely related to that of Linzen et al. (2016) and Bernardy and Lappin (2017), which we discussed in the introduction. Other recent work has focused on the morphological and grammatical knowledge that RNN-based machine-translation systems and sentence embeddings encode, typically by training classifiers to decode various linguistic properties from hidden states of the network (e.g., Adi et al., 2017; Belinkov et al., 2017; Shi et al., 2016), or looking at whether the end-to-end system correctly translates sentences with challenging constructions (Sennrich, 2017).

Previous work in neurolinguistics and psycholinguistics used jabberwocky, or pseudo-word, sentences to probe how speakers process syntactic information (Friederici et al., 2000; Moro et al.,

[12] These subject-verb agreement results are in general higher than for our own subject-verb agreement construction (NOUN VERB VERB) because the latter always includes an embedded clause, and it is therefore harder on average.

[13] Similarly high performance of LM-trained RNNs on Linzen's dataset was recently reported by Yogatama et al. (2018).

2001; Johnson and Goldberg, 2013). Such sentences are obtained by substituting original words with morphologically and phonologically acceptable nonce forms. We are not aware of work that used nonce sentences made of real words to evaluate the syntactic abilities of models or human subjects. As a proof of concept, Pereira (2000) and, later, Mikolov (2012) computed the probability of Chomsky's famous "colorless green ideas" sentence using a class-based bigram LM and an RNN, respectively, and showed that it is much higher than the probability of its shuffled ungrammatical variants.

6 Conclusion

We ran an extensive analysis of the abilities of RNNs trained on a generic language-modeling task to predict long-distance number agreement. Results were consistent across four languages and a number of constructions. They were above strong baselines even in the challenging case of nonsense sentences, and not far from human performance. We are not aware of other collections of human long-distance agreement judgments on nonsensical sentences, and we thus consider our publicly available data set an important contribution of our work, of interest to students of human language processing in general.

The constructions we considered are quite infrequent (according to a rough estimate based on the treebanks, the language in which they are most common is Hebrew, and even there they occur with average 0.8% sentence frequency). Moreover, they vary in the contexts that separate the cue and the target. So, RNNs are not simply memorizing frequent morphosyntactic sequences (which would already be impressive, for systems learning from raw text). We tentatively conclude that LM-trained RNNs can construct abstract grammatical representations of their input. This, in turn, suggests that the input itself contains enough information to trigger some form of syntactic learning in a system, such as an RNN, that does not contain an explicit prior bias in favour of syntactic structures.

In future work, we would like to better understand what kind of syntactic information RNNs are encoding, and how. On the one hand, we plan to adapt methods to inspect information flow across RNN states (e.g., Hupkes et al., 2017). On the other, we would like to expand our empirical investigation by focusing on other long-distance phenomena, such as overt case assignment (Blake, 2001) or parasitic gap licensing (Culicover and Postal, 2001). While it is more challenging to extract reliable examples of such phenomena from corpora, their study would probe more sophisticated syntactic capabilities, possibly even shedding light on the theoretical analysis of the underlying linguistic structures. Finally, it may be useful to complement the corpus-driven approach used in the current paper with constructed evaluation sentences that isolate particular syntactic phenomena, independent of their frequency in a natural corpus, as is common in psycholinguistics (Enguehard et al., 2017).

Acknowledgements

We thank the reviewers, Germán Kruszewski, Gerhard Jäger, Adam Liška, Tomas Mikolov, Gemma Boleda, Brian Dillon, Cristophe Pallier, Roberto Zamparelli and the Paris Syntax and Semantics Colloquium audience for feedback and advice.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR Conference Track*. Toulon, France. Published online: https://openreview.net/group?id=ICLR.cc/2017/conference.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of ACL*. Vancouver, Canada, pages 861–872.

Jean-Philippe Bernardy and Shalom Lappin. 2017. Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology* 15(2):1–15.

Barry Blake. 2001. *Case*. MIT Press, Cambridge, MA.

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

Bo Cartling. 2008. On the implicit acquisition of a context-free grammar by a simple recurrent neural network. *Neurocomputing* 71:1527–1537.

Ciprian Chelba, Mohammad Norouzi, and Samy Bengio. 2017. N-gram language modeling using recurrent neural network estimation. *arXiv preprint arXiv:1703.10724* .

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, Berlin, Germany.

Morten Christiansen and Nick Chater. 1999. Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science* 23(2):157–205.

James Cross and Liang Huang. 2016. Incremental parsing with minimal features using bi-directional LSTM. In *Proceedings of ACL (Short Papers)*. Berlin, Germany, pages 32–37.

Peter Culicover and Paul Postal, editors. 2001. *Parasitic gaps*. MIT Press, Cambridge, MA.

Jeffrey Elman. 1990. Finding structure in time. *Cognitive Science* 14:179–211.

Jeffrey Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7:195–225.

Jeffrey Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition* 48:71–99.

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.

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

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. Irstlm: An open source toolkit for handling large scale language models. In *Ninth Annual Conference of the International Speech Communication Association*. pages 1618–1621.

Angela D. Friederici, Martin Meyer, and D. Yves Von Cramon. 2000. Auditory language comprehension: An event-related fMRI study on the processing of syntactic and lexical information. *Brain and Language* 74(2):289–300.

Felix Gers and Jürgen Schmidhuber. 2001. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks* 12(6):1333–1340.

Alex Graves. 2012. *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, Berlin.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 187–197.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–178–.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2017. Visualisation and diagnostic classifiers reveal how recurrent and recursive neural networks process hierarchical structure. http://arxiv.org/abs/1711.10203.

Matt A. Johnson and Adele E. Goldberg. 2013. Evidence for automatic accessing of constructional meaning: Jabberwocky sentences prime associated verbs. *Language and Cognitive Processes* 28(10):1439–1452.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics* 4:313–327.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*. IEEE, volume 1, pages 181–184.

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.

Heidi Lorimor, Kathryn Bock, Ekaterina Zalkind, Alina Sheyman, and Robert Beard. 2008. Agreement and attraction in Russian. *Language and Cognitive Processes* 23(6):769–799.

Tomas Mikolov. 2012. *Statistical language models based on neural networks*. Dissertation, Brno University of Technology.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of INTERSPEECH*. Makuhari, Japan, pages 1045–1048.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL*. Atlanta, Georgia, pages 746–751.

Andrea Moro, Marco Tettamanti, Daniela Perani, Caterina Donati, Stefano F Cappa, and Ferruccio Fazio. 2001. Syntax and the brain: disentangling grammar by selective anomalies. *Neuroimage* 13(1):110–118.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, 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 Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno,

Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Fernando Pereira. 2000. Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 358(1769):1239–1253.

Douglas Rohde and David Plaut. 1997. Simple recurrent networks and natural language: How important is starting small? In *Proceedings of CogSci*. Stanford, CA, pages 656–661.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL-SIGDAT Workshop*. Dublin, Ireland.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of EACL (Short Papers)*. Valencia, Spain, pages 376–382.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of EMNLP*. Austin, Texas, pages 1526–1534.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. http://arxiv.org/abs/1609.08144.

Dani Yogatama, Yishu Miao, Gabor Melis, Wang Ling, Adhiguna Kuncoro, Chris Dyer, and Phil Blunsom. 2018. Memory architectures in recurrent neural network language models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkFqf0lAZ.

Xingxing Zhang, Jianpeng Cheng, and Mirella Lapata. 2017. Dependency parsing as head selection. In *Proceedings of EACL*. Valencia, Spain, pages 665–676.