Colorless green recurrent networks dream hierarchically

Kristina Gulordava[*]
Universitat Pompeu Fabra

Piotr Bojanowski
Facebook AI Research

Edouard Grave
Facebook AI Research

Tal Linzen
Johns Hopkins University

Marco Baroni
Facebook AI Research

Introduction. Recurrent neural networks (RNNs) are general sequence processing devices that do not explicitly encode hierarchical structure. Early work using artificial languages showed that they may nevertheless be able to approximate context-free languages (Elman, 1991). 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, 2012; Graves, 2012). These developments re-fueled the old debate between generative linguistics and connectionism on whether neural models can represent syntax without explicitly encoding hierarchical structure (Pater, 2017).

Linzen et al. (2016) evaluated the extent to which RNNs can approximate hierarchical structure by testing whether these models can learn English subject-verb agreement, a task thought to require hierarchical structure in the general case ("the girl the boys like…" is or are?). In their study, RNNs could only succeed when provided with explicit supervision on the target task. RNNs that were instead trained to perform generic, large-scale language modeling (predicting the next word given the context), with no explicit focus on agreement at training time, did not pass the test.

The current work re-evaluates the conclusions of Linzen et al. (2016), showing that RNNs trained on generic language modeling can successfully predict agreement in hard cases. We strengthen the evaluation paradigm of Linzen and colleagues in several ways. Most importantly, 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 also test long-distance agreement in sentences that are grammatical but completely meaningless, e.g., (paraphrasing Chomsky): "The colorless green ideas I ate with the chair sleep furiously". We thus evaluate a stricter form of grammaticality compared to the broadly studied notion of acceptability (which is affected by semantics and discourse). "Colorless green" sentences should be particularly hard for statistical models since they do not contain surface features (e.g., ngrams) observed during training. Indeed, Chomsky argues that such models cannot judge the grammaticality of nonce sentences. To our knowledge, we present the first systematic experiments to bear on this claim.

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 a new public repository of grammaticality data and an important comparison point for RNN performance.

Data. We automatically extracted number agreement constructions from the Universal Dependencies treebanks.[1] We collected cue-target pairs of categories connected by a dependency relation in the treebank, and which displayed matching number features (e.g., subject noun and main verb). To extract hard constructions where agreement cannot be predicted from linear adjacency, we collected only instances where cue and target were separated by at least three tokens. This step resulted in between two (English, the language with the poorest morphology) and 21 (Russian) constructions per language, and between 41 and 442 sentences.

Based on these corpus-extracted instances, we constructed a dataset of "colorless green" nonce sentences. We generated nine nonce variants of each original sentence. Each content word (noun,

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

[1] http://universaldependencies.org/

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 363-364.
New York City, New York, January 3-6, 2019

	Italian	English	Hebrew	Russian
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

Table 1: Accuracy averaged across the five best models in terms of perplexity on the validation set.

verb, adjective, proper noun, numeral, adverb) in the sentence was substituted by another random content word from the treebank with matching category and morphological features. Function words (determiners, pronouns, adpositions, particles) and punctuation were left intact. For example, we generated the nonce (1b) from the original conjoined-verb-agreement sentence (1a):

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

Model. We trained long-short term memory RNNs (LSTMs, Hochreiter and Schmidhuber, 1997) on 90M tokens of Wikipedia text in each language, using 10M more tokens for validation. The models were trained and tuned on language modeling, and *not* on predicting agreement.

Evaluation. 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), the model should assign a higher probability to "finds" than "find".

Results. Our main result is that RNNs trained on generic language modeling handle long-distance agreement well, even on nonce sentences, and consistently across languages (Table 1). They achieve high accuracy well above baselines based on surface ngrams (not reported here).

To challenge Chomsky's claim that statistical models cannot mimic speakers' competence in assessing the grammaticality of "colorless green" sentences, we compared model and human performance in Italian. In a Mechanical Turk experiment, subjects were requested to finish the sentences in our test set by choosing between singular and plural forms. The average human accuracy was 94.5 and 88.4% on original and nonce sentences respectively. Similarly to the model results, there was a consistent gap in human accuracy between original and nonce sentences (6.1%). Crucially, the gap in accuracy between the human subjects and the model was relatively 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 vs. 92.3%).

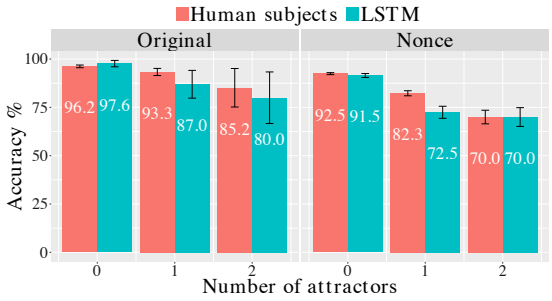

Figure 1: Accuracy by number of attractors.

Attractors, that is, intermediate words with the same category of the cue but opposite number, constitute an obvious challenge for agreement processing (Bock and Miller, 1991). We show how their presence affects human and model behavior in Fig. 1. 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 LSTM RNN are quite robust to the presence of attractors, in contrast to what was reported by Linzen et al. (2016).

References

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

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

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

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

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1787.

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

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

J. Pater. 2017. Generative linguistics and neural networks at 60: foundation, friction, and fusion .