Evaluating recurrent neural networks as cognitive models of syntax

Tal Linzen[1,2], Émile Enguehard[1], Yoav Goldberg[3] and Emmanuel Dupoux[1]
[1]LSCP & IJN, CNRS, EHESS, ENS, PSL Research University
[2]Department of Cognitive Science, Johns Hopkins University
[3]Computer Science Department, Bar Ilan University

Abstract

We analyze the ability of modern recurrent neural networks (RNNs) to learn subject-verb number agreement, a phenomenon considered to require a representation of the structure of the sentence; our goal was to assess the potential of RNNs to serve as cognitive models of syntax. Given explicit supervision, the network made few errors overall, but error rates increased in more complex sentences, and the pattern of errors was qualitatively different than in human experiments. Multi-task learning narrowed the gap between RNNs and humans. In summary, RNNs may be able to develop structural representations, but require strong supervision to do so.

Keywords: Recurrent neural networks, psycholinguistics, syntax, agreement

Large-scale artificial neural networks have reportedly begun to approach human-level performance in some tasks. A deeper understanding of the performance of these networks could shed light on how the human brain might accomplish those tasks. Yet the potential of neural networks for cognitive science is limited by the fact that available evaluations of neural network focus on applied engineering goals rather than clearly defined cognitive capacities. We begin to address this gap in the domain of language; this domain is particularly fruitful for this purpose since we can draw on the precise characterization of the human knowledge of language provided by the linguistics literature. We focus on a single syntactic phenomenon, English subject-verb agreement; as we illustrate below, to fully master this construction, a learner needs to develop a sophisticated understanding of sentence structure.

Subject-verb agreement

English present-tense third-person verbs agree in number with their subject: singular subjects require singular verbs (*the boy smiles*) and plural subjects require plural verbs (*the boys smile*). Identifying the subject of a particular verb can be non-trivial in sentences with multiple nouns:

(1) The only championship **banners** that are currently displayed within the building **are** for national or NCAA Championships.

Determining that the subject of the verb in boldface is *banners* rather than the singular nouns *championship* and *building* requires an understanding of the structure of the sentence.

We test the ability of an architecure to learn this phenomenon using the agreement task (Bock & Miller, 1991; Elman, 1991). In this task, the learner is given the words leading up to a verb (a "preamble"), and is instructed to predict whether that verb will take the plural or singular form. In simple cases, this task can be solved using heuristics such as copying the number of the most recent noun. Since such heuristics are of limited cognitive interest, we focus our evaluation on more difficult sentences in which one or more nouns of the opposite number from the subject intervene between the subject and the verb; such nouns are termed "attractors" since they "attract" the agreement away from the subject.

Agreement processing in RNNs

We trained recurrent neural networks (RNNs) with 50 Long Short-Term Memory (LSTM) units (Hochreiter & Schmidhuber, 1997) in one of two regimes. In the *number prediction* regime, we gave it specific instruction on the task: a large number of preambles along with the correct number of the verb following each preamble. In the *language modeling* regime, we instructed it to predict a full probability distribution over the next word at each point in the sentence; in the evaluation stage, the task was performed by comparing the probability assigned by the network to the grammatically appropriate and inappropriate forms of the verb. We trained all networks on approximately 121000 sentences from the English Wikipedia, and evaluated them on new sentences from the same corpus.

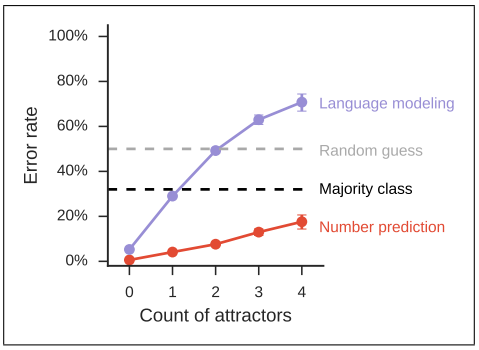

Figure 1: Error rate as a function of the count of attractors.

The overall error rate in the number prediction regime was very low (0.83%). Error rates increased on sentences with attractors, indicating that the RNNs did not fully master the phenomenon (Figure 1); at the same time, even the 18% error rate on sentences with four attractors was substantially better than the last-noun heuristic (100%), random choice (50%) or the majority class (32% for always predicting a "singular" verb). The effect of attractors was much more dramatic in the language modeling regime: error rates jumped to 30% with one attractor, and were worse than random guessing for three

and four attractors. An analysis of the number prediction networks' activation showed that some units specialized to track interpretable properties of the sentence, such as the number of the main clause subject or the embedded clause subject. We conclude that RNNs can learn to approximate sentence structure sufficiently to inflect verbs with high accuracy, but only when their attention is directed to the grammatical task.

Comparison to human errors

Humans occasionally make agreement errors, such as *the time for fun and games are over*. This phenomenon has been studied in controlled experiments, starting with Bock and Miller (1991). Three relevant findings of these experiments can be summarized as follows: (a) errors are much more likely when there is an attractor; (b) errors are more likely when the subject is singular and the attractor is plural than the other way around; (c) errors are equally likely when the modifier of the subject is a relative clause or a prepositional phrase, even though relative clauses are arguably more complex:

(2) Relative: The **tape** that promoted the rock singers...

(3) Prepositional: The **tape** from the rock singers...

The model replicated finding (a) – overall error rates for sentences with an attractor were around 6% (Figure 1), somewhat less than the proportion of errors that humans make in speeded production (Bock & Miller, 1991), around 10%. To test (b) and (c), we used the number prediction networks described above to predict the number of the verb in the materials of Bock and Cutting (1992), which follow the relative clause and prepositional phrase patterns shown above. The networks did not show a consistent asymmetry between singular-plural and plural-singular sentences (see the single-task agreement model in Figure 2). Errors were dramatically more frequent in relative clauses than in prepositional phrases. In summary, the generally good performance of the RNNs on sentences from the corpus contrasts with qualitative differences in error patterns between the RNNs and humans: the RNNs represent complex sentences much less accurately than humans do.

Inductive bias from multi-task learning

The supervision that humans receive is likely to be richer than the supervision we gave to our RNNs; for example, visual scenes or action plans may have hierarchical structure that could serve as an inductive bias for language learning and thereby improve the acquisition of complex sentences. Are the limitations that RNNs showed in the previous sections inherent to their architecture, or can these limitations be mitigated by stronger supervision? We address this question using multi-task learning, where the same model is encouraged to develop representations that are simultaneously useful for multiple tasks. We first trained the RNNs to perform "CCG supertagging", a standard syntactic natural language processing task; the resulting RNN weights were then used to initialize training on the agreement task as above.

As shown in Figure 2, this training regime improved the performance of the RNNs on relative clauses, though still did not bring it to human levels. A suggestive asymmetry emerged between singular and plural subjects. While CCG supertagging is not a direct proxy to any particular form of supervision that humans may receive, this experiment shows that RNNs can indeed develop improved syntactic representations without explicit changes to their architecture.

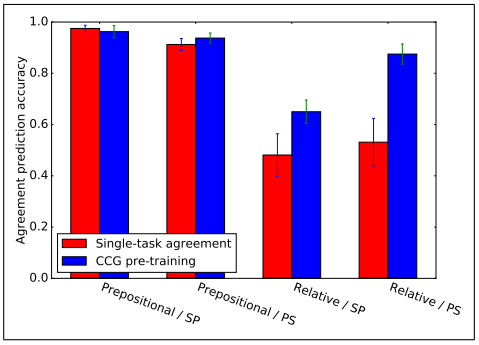

Figure 2: Comparison between prepositional phrase and relative clauses modifiers (Bock & Cutting, 1992). SP indicates a singular subject and a plural attractor.

Conclusion

Our experiments show that RNNs are able to approximate sentence structure in most cases, but falter in complex sentences in ways that are qualitatively different from humans. The degree of supervision played a crucial role in the quality of the representations. More generally, our results show that the detailed characterization of human-level knowledge of language produced by linguists can help us understand the strengths and weaknesses of existing architectures and analyze models that are otherwise difficult to interpret.

Acknowledgments

This research was supported by the European Research Council (grant ERC-2011-AdG 295810 BOOTPHON), the Agence Nationale pour la Recherche (grants ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC) and the Israeli Science Foundation (grant number 1555/15).

References

Bock, K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, *31*(1), 99–127.

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*(2-3), 195–225.

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