Large neural network language models learn representations of incremental parse states

Richard Futrell[1], Ethan Wilcox[2], Takashi Morita[3], Peng Qian[4], Miguel Ballesteros[5], Roger Levy[4]
[1]UC Irvine, [2]Harvard University, [3]Kyoto University, [4]MIT, [5]IBM
Contact: rfutrell@uci.edu

The modern field of natural language processing is dominated by neural network methods, in which a generic sequence model such as a Recurrent Neural Network (RNN: Elman, 1990; Hochreiter & Schmidhuber, 1997) is used to convert a sentence into a vector-space representation (Sutskever et al., 2014; Wu et al., 2016; Peters et al., 2018). However, the sentence representations derived in this way are opaque to human interpretation: in particular, it is not known what information the vector representation of a sentence contains about about the syntactic parse tree of the sentence. We treat neural networks as black boxes and perform behavioral experiments on them to elicit information about their syntactic representations as if they were human subjects in psycholinguistics studies. We show that RNNs maintain representations of parse states as evidenced by garden path effects. Our work shows that these generic models can learn at least coarse-grained representations of syntactic structure from distributional evidence. We join work showing strong parallels between RNNs and human sentence processing (Christiansen & Chater, 1999; MacDonald & Christiansen, 2002; Frank & Bod, 2011; van Schijndel & Linzen, 2018), but we also find some qualitative differences between RNN behavior and human sentence processing.

We study RNN **language models**: models which, given a string prefix, produce a probability distribution over continuations. We use garden path sentences as stimuli, to ask whether the continuation distributions reflect the same kind of behavior which would be captured using a stack-like parse representation in a system such as an incremental PCFG parser (Stolcke, 1995). We use the neural network's **surprisal** at a word ($-\log p(word|context)$) as an analogue of word-by-word reading times (Hale, 2001; Levy, 2008; Smith & Levy, 2013): high surprisal at a disambiguating word indicates a garden path effect because the word was not expected under the network's prefered parse of the locally ambiguous material. We study two publically available RNN language models: "**JRNN**" (Jozefowicz et al., 2016), trained on nearly one billion tokens of text, equivalent to a human lifetime of linguistic input; and "**GRNN**" (Gulordava et al., 2018), trained on 100 million tokens, equivalent to six years of linguistic input.

We demonstrate three phenomena in the RNN language models: (1) Main Verb/Reduced-Relative (MV/RR) garden path effects modulated by verb-form ambiguity (see Fig. 1); (2) MV/RR garden path effects modulated by head noun animacy (Trueswell et al., 1994; see Fig. 2); and (3) NP/Z garden path effects modulated by the transitivity of the embedded verb (Staub, 2007; see Fig. 3). We find that all language models show garden path effects indicating representation of parse state, but the networks are inconsistent in their use of fine-grained lexical cues signalling the beginnings and endings of such states. We also study whether RNN language models show **digging-in effects** in the NP/Z ambiguity (Tabor & Hutchins, 2004; Levy et al, 2009), in which garden path effects become larger as the distance between the onset of the local ambiguity and the disambiguator is increased. Instead we find that RNNs show *reverse* digging-in effects: garden path effects become *smaller* with increasing the distance between the onset of the local ambiguity and the disambiguator (see Fig. 4).

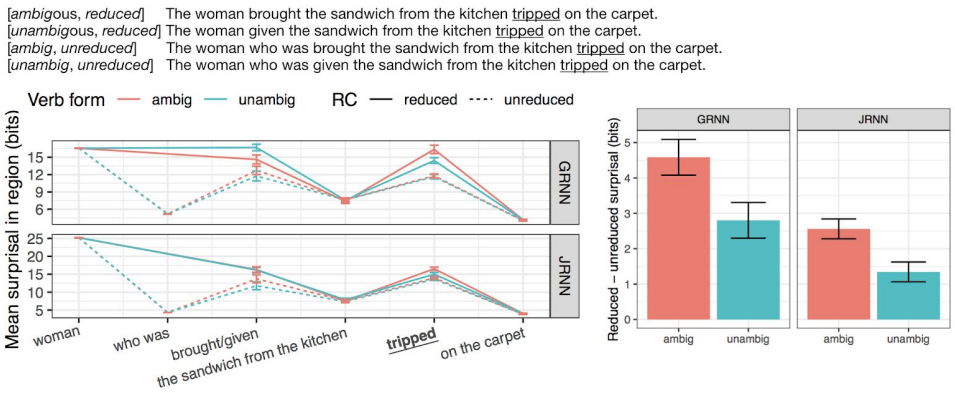

Fig 1. Top: example stimuli for MV/RR garden path modulated by verb-form ambiguity. Disambiguating region is underlined. **Bottom left**: Region-by-region surprisal values. **Bottom right**: Average surprisal differences (*reduced* condition minus *unreduced* condition) at the disambiguating region *tripped*. Error bars are 95% confidence intervals of the contrasts between conditions (Masson & Loftus, 2003). *n*=28 items. Both GRNN and JRNN show a significant interaction of verb-form ambiguity and reduction.

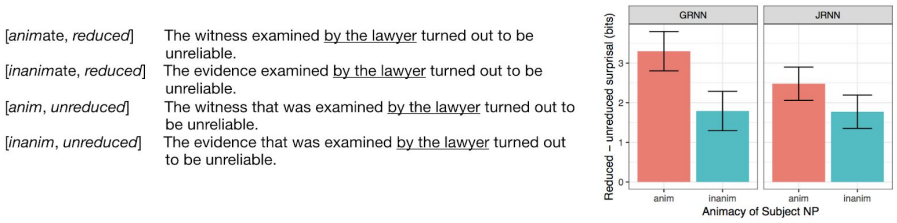

Fig 2. Left: example stimuli for MV/RR garden path modulated by subject animacy. **Right**: Average surprisal differences (*reduced* condition minus *unreduced* condition) at the disambiguating region *by the lawyer*. *n*=30 items. Only JRNN shows a significant interaction of subject animacy and reduction.

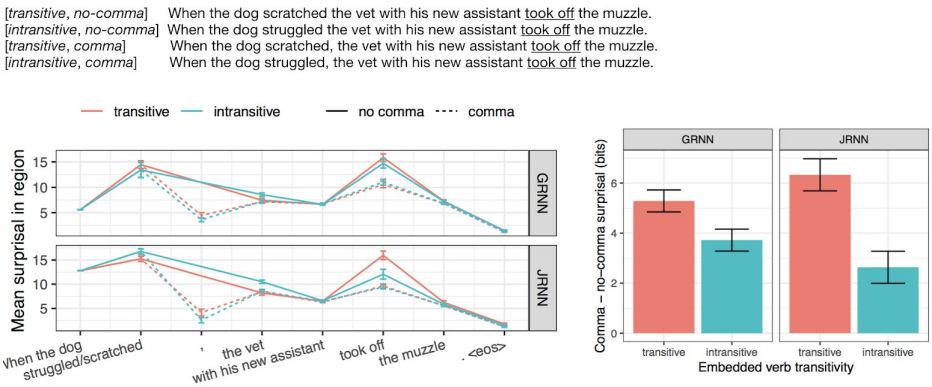

Fig 3. Top: example stimuli for NP/Z garden path modulated by embedded verb transitivity. **Bottom left**: Region-by-region surprisal values. **Bottom right**: Average surprisal values at the disambiguating region *took off*. *n*=24 Items. Only JRNN show a significant interaction of transitivity and comma presence.

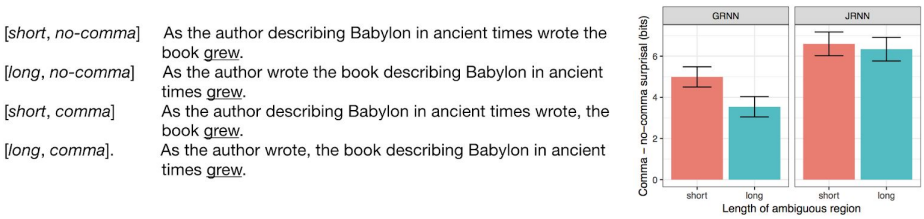

Fig 4. Left: example stimuli for NP/Z garden path with length manipulation. **Right**: Average surprisal values at the disambiguating region *grew*. *n*=31 items. GRNN has a significant negative interaction of length with comma presence.