What do Recurrent Neural Network Grammars Learn About Syntax ?

Authors: Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, Noah A. Smith

Presented by: Triveni Putti

Paper link: https://arxiv.org/pdf/1611.05774.pdf

Contents

- Recap of Recurrent Neural Network Grammars (RNNGs)
- Outline of the paper
- Ablated RNNGs
  - Experiments and results
- Gated Attention RNNGs
  - Experiments and results
  - Headedness in phrases
- Role of Non-Terminal Labels
- Key Takeaways

RNNGs

- Language is hierarchical
- Generate **symbols** sequentially using an **RNN**
- Add some **control symbols** to rewrite the history occasionally
  - Occasionally **compress** a sequence into a constituent
  - RNN predicts next terminal/control symbol based on the history of compressed elements and non-compressed terminals

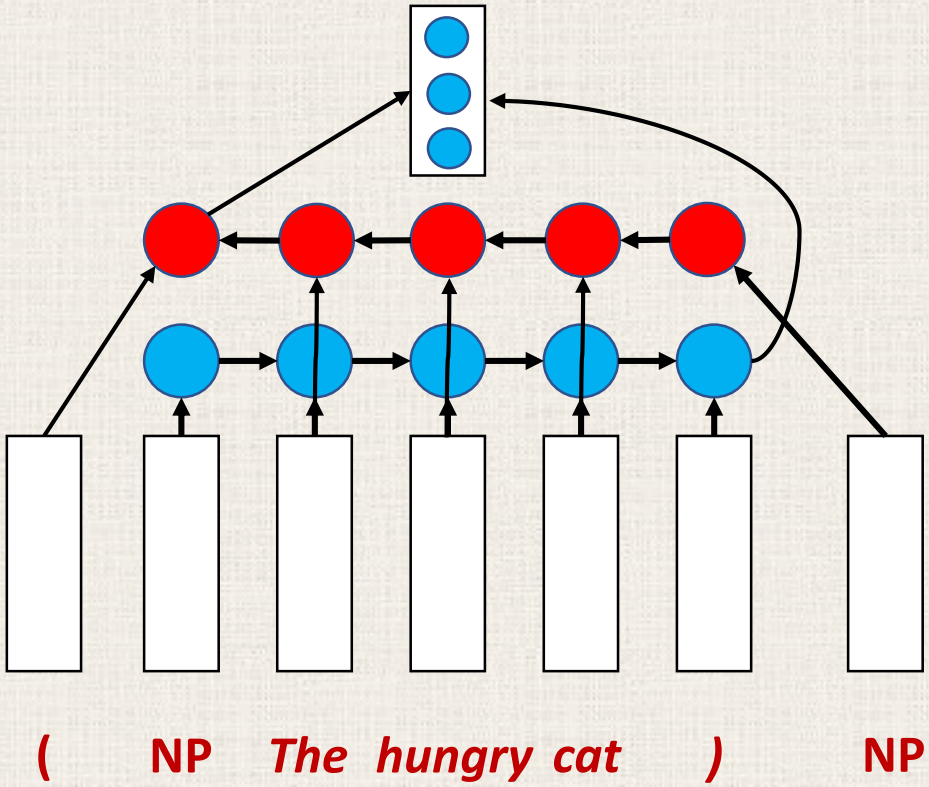
Example

The hungry cat meows.

Terminals	Stack	Action
		NT(S)
	(S	NT(NP)
	(S (NP	GEN(*The*)
The	(S (NP *The*	GEN(*hungry*)
The hungry	(S (NP *The hungry*	GEN(*cat*)
The hungry cat	(S (NP *The hungry cat*	REDUCE
The hungry cat	(S (NP *The hungry cat*)	NT(VP)
The hungry cat	(S (NP *The hungry cat*) (VP	GEN(*meows*)
The hungry cat meows	(S (NP *The hungry cat*) (VP *meows*	REDUCE
The hungry cat meows	(S (NP *The hungry cat*) (VP meows)	GEN(.)
The hungry cat meows.	(S (NP *The hungry cat*) (VP meows).	REDUCE
The hungry cat meows.	(S (NP *The hungry cat*) (VP meows).)	

Composition Function

Bidirectional LSTM used for the representation for: (NP *The hungry cat*)



What is the paper about ?

- What is the information RNNGs exactly learn from a linguistic perspective?

- Approach:
  1. Modify the models to discover the importance of composition function
  2. Augment the composition function with gated attention mechanism (leading to GA-RNNG)
     - Role that individual heads play in phrasal representation
     - Role that non terminal labels play

Composition Function is key

Model	F_1
Vinyals et al. (2015) – PTB only	88.3
Discriminative RNNG	**91.2**
Choe and Charniak (2016) – PTB only	92.6
Generative RNNG	**93.3**

Exp. 1: Phrase structure parsing performance on PTB

- Both discriminative and generative RNNGs have higher accuracy for phrase structure parsing
- RNNGs explicit composition function which the other two models must learn implicitly plays a key role.

Ablated RNNGs

- All the three data structures – stack, buffer and action history are redundant. For instance, every generated word (stored in buffer) goes into stack too.
- But stack only has the composition function and not the other two. So, we expect that only stack is critical to the RNNG's performance.
- To test this conjecture, experiments were carried out on ablated RNNGs that lack each of the 3 data structures, and one that lacks both action history and buffer.

What do we expect ?

Ablated RNNGs - Results

Model	F_1
Vinyals et al. (2015)†	92.1
Choe and Charniak (2016)	92.6
Choe and Charniak (2016)†	**93.8**
Baseline RNNG	93.3
Ablated RNNG (no history)	93.2
Ablated RNNG (no buffer)	93.3
Ablated RNNG (no stack)	92.5
Stack-only RNNG	**93.6**

Exp. 2 : Phase structure parsing performance on PTB

1. Stack- only RNNG is the best among supervised models and even outperforms the full RNNG

2. Ablating the stack gives worst performance (supports the importance of composition)

+ indicates systems that use additional unparsed data (semi supervised)

Ablated RNNGs - Results

Model	UAS	LAS
Kiperwasser and Goldberg (2016)	93.9	91.9
Andor et al. (2016)	94.6	92.8
Dozat and Manning (2016)	95.4	93.8
Choe and Charniak (2016)†	**95.9**	94.1
Baseline RNNG	95.6	94.4
Ablated RNNG (no history)	95.4	94.2
Ablated RNNG (no buffer)	95.6	94.4
Ablated RNNG (no stack)	95.1	93.8
Stack-only RNNG	**95.8**	**94.6**

Exp. 3: Dependency parsing performance on PTB

1. Stack- only RNNG is the best among supervised models and even outperforms the full RNNG

2. Ablating the stack gives worst performance (supports the importance of composition)

Ablated RNNGs - Results

Model	Test ppl. (PTB)
IKN 5-gram	169.3
LSTM LM	113.4
RNNG	105.2
Ablated RNNG (no history)	105.7
Ablated RNNG (no buffer)	106.1
Ablated RNNG (no stack)	113.1
Stack-only RNNG	**101.2**

Exp. 4 : Language modeling : Perplexity

1. Stack- only RNNG is the best among supervised models and even outperforms the full RNNG

2. Ablating the stack gives worst performance (supports the importance of composition)

Gated Attention RNNG - Understanding the learnt phrasal representations

- Having established that the composition function is key to RNNG's performance, let's see the nature of composed phrasal representations
- Interpreting the composition function for most NNs is difficult.
- Fortunately, we have some hypotheses offered by linguistic theories about the nature of representation of phrases
- Two of such hypotheses are looked at in this paper:
  - Phrasal representations are strongly determined by an individual/ multiple lexical head(s).
  - The representations combine all children without any salient head

Gated Attention Composition

- Variant of the composition function that uses explicit attention mechanism and a sigmoid gate with multiplicative interactions
- Assign an “attention weight” to each of the children. Parent phrase is represented by the combination of sum of each child’s representation scaled by its attention weight and its nonterminal type.

$$\mathbf{g} = \sigma\left(\mathbf{W}_1 \mathbf{t}_{nt} + \mathbf{W}_2 \mathbf{m} + \mathbf{b}\right)$$

- The final phrasal representation is an element wise multiplication w.r.t. t_{nt} and m

$$\mathbf{c} = \mathbf{g} \odot \mathbf{t}_{nt} + (1 - \mathbf{g}) \odot \mathbf{m}.$$

Gated Attention RNNG- Results

Exp. 2 : Phase structure parsing performance on PTB

Model	F_1
Vinyals et al. (2015)†	92.1
Choe and Charniak (2016)	92.6
Choe and Charniak (2016)†	**93.8**
Baseline RNNG	93.3
Ablated RNNG (no history)	93.2
Ablated RNNG (no buffer)	93.3
Ablated RNNG (no stack)	92.5
Stack-only RNNG	**93.6**
GA-RNNG	93.5

Exp. 3: Dependency parsing performance on PTB

Model	UAS	LAS
Kiperwasser and Goldberg (2016)	93.9	91.9
Andor et al. (2016)	94.6	92.8
Dozat and Manning (2016)	95.4	93.8
Choe and Charniak (2016)†	**95.9**	94.1
Baseline RNNG	95.6	94.4
Ablated RNNG (no history)	95.4	94.2
Ablated RNNG (no buffer)	95.6	94.4
Ablated RNNG (no stack)	95.1	93.8
Stack-only RNNG	**95.8**	**94.6**
GA-RNNG	95.7	94.5

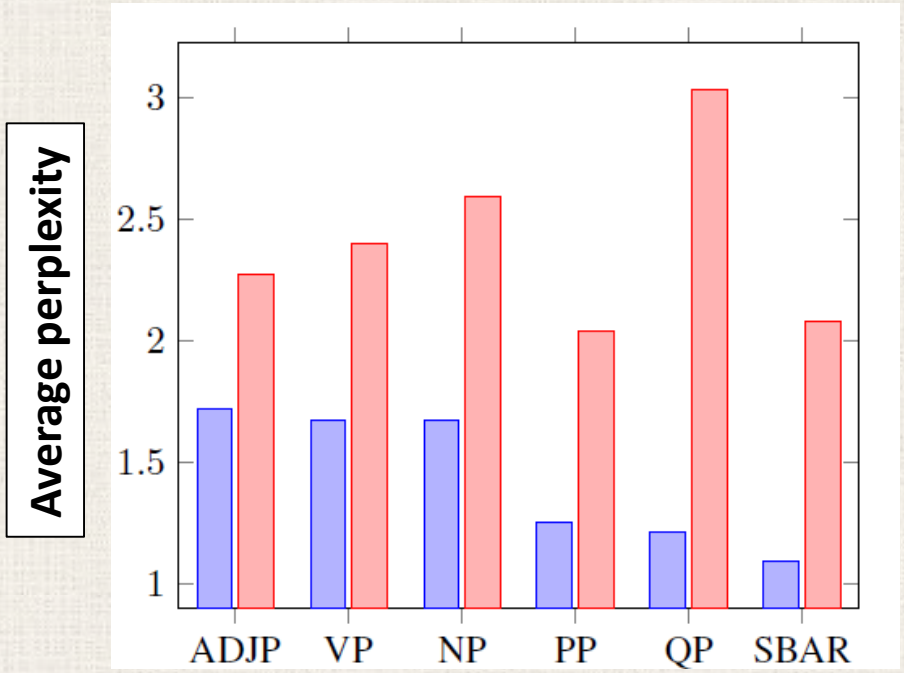
Exp. 4 : Language modeling : Perplexity

Model	Test ppl. (PTB)
IKN 5-gram	169.3
LSTM LM	113.4
RNNG	105.2
Ablated RNNG (no history)	105.7
Ablated RNNG (no buffer)	106.1
Ablated RNNG (no stack)	113.1
Stack-only RNNG	**101.2**
GA-RNNG	**100.9**

Gated RNNG outperforms Baseline RNNG and achieves competitive performance with stack-only variant

Headedness

- Attention weights can tell us which constituents are most important to a phrase's vector representation in the stack
- Headedness is centering the attention around a single or few elements



- Average perplexity can be interpreted as the average number of "choices" for each nonterminal category
- Blue represents the learned attention vectors on the test set and red represents the uniform distribution (no headedness)
- Since the weights have much lower perplexity than the uniform distribution baseline, they are quite peaked around certain components.

Headedness- Distribution for major NTs

	Prepositional phrases
1	ADVP (0.14) **on (0.72)** NP (0.14)
2	ADVP (0.05) **for (0.54)** NP (0.40)
3	ADVP (0.02) **because (0.73)** of (0.18) NP (0.07)
4	such (0.31) **as (0.65)** NP (0.04)
5	from (0.39) **NP (0.49)** PP (0.12)
6	**of (0.97)** NP (0.03)
7	**In (0.93)** NP (0.07)
8	**by (0.96)** S (0.04)
9	**at (0.99)** NP (0.01)
10	NP (0.1) **after (0.83)** NP (0.06)

Attention weight vectors for some samples for PPs

In almost all the examples, prepositions are given the most attention

Headedness- Distribution for major NTs

	Noun phrases
1	Canadian (0.09) **Auto (0.31)** Workers (0.2) union (0.22) president (0.18)
2	no (0.29) major (0.05) **Eurobond (0.32)** or (0.01) foreign (0.01) bond (0.1) offerings (0.22)
3	Saatchi (0.12) client (0.14) Philips (0.21) Lighting (0.24) **Co. (0.29)**
4	nonperforming (0.18) commercial (0.23) **real (0.25)** estate (0.1) **assets (0.25)**
5	the (0.1) Jamaica (0.1) Tourist (0.03) Board (0.17) ad (0.20) **account (0.40)**
6	the (0.0) final (0.18) **hour (0.81)**
7	their (0.0) first (0.23) **test (0.77)**
8	**Apple (0.62)** , (0.02) Compaq (0.1) and (0.01) IBM (0.25)
9	both (0.02) stocks (0.03) and (0.06) **futures (0.88)**
10	NP (0.01) , (0.0) **and (0.98)** NP (0.01)

Attention weight vectors for some samples for NPs

Simple NPs – Rightmost nouns> Adjectives> Determiners ~ Possessive determiners(6,7)

Complex NPs – Both first (8) or last noun (9) can have high attention; for conjunctions of multiple NPs, conjunction gets most attention (10)

Headedness- Distribution for major NTs

	Verb phrases
1	buying (0.31) and (0.25) selling (0.21) NP (0.23)
2	ADVP (0.27) show (0.29) PRT (0.23) PP (0.21)
3	pleaded (0.48) ADJP (0.23) PP (0.15) PP (0.08) PP (0.06)
4	received (0.33) PP (0.18) NP (0.32) PP (0.17)
5	cut (0.27) NP (0.37) PP (0.22) PP (0.14)
6	to (0.99) VP (0.01)
7	were (0.77) n't (0.22) VP (0.01)
8	did (0.39) n't (0.60) VP (0.01)
9	handle (0.09) NP (0.91)
10	VP (0.15) and (0.83) VP 0.02)

Attention weight vectors for some samples for VPs

Simple VPs - NP> Verb (9); Negation is assigned non-trivial weight (7,8)

Other VPs - for conjunctions of multiple VPs, conjunction gets most attention (10)

Headedness - Comparison to Existing Head rules

- Overlap is measured between the above results and two head rules : Collins and Stanford
- Model has higher overlap with the Collins head rules rather than the Stanford
- This can be attributed to the fact that Stanford incorporates semantic considerations while RNNG is purely syntactical
- The major disagreement is with the attention weight in a VP (attention is given to NP instead of Verb)

GA-RNNG can infer head rules to a large extent.

Role of Non-Terminal Labels

- Are heads sufficient to create representations of phrases or whether extra nonterminal information is necessary?
- GA-RNNG is trained on unlabeled trees (only bracketings without nonterminal types) denoted as U-GA-RNNG
- On test data, the GA-RNNG achieves 94.2% parsing accuracy, while the U-GA-RNNG achieves 93.5%
- This result suggests that nonterminal labels add a relatively small amount of information and bracketings are the most important part

Conclusion

1. The composition function, a key differentiator between the RNNG and other neural models of syntax, is crucial for good performance.

2. Using the attention vectors we discover that the model is learning something similar to heads, although the attention vectors are not completely peaked around a single component.

3. Bracketing annotation does most of the work of syntax making phrasal representations depend minimally on non-terminals.

QUESTIONS ?