Deep RNNs Encode Soft Hierarchical Syntax

Terra Blevins, Omer Levy, and Luke Zettlemoyer
Paul G. Allen School of Computer Science & Engineering
University of Washington, Seattle, WA

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (2018)

Presented by

Tim Patzelt

in

Recent Highlights in NLP, University of Potsdam

24.06.2019

Outline

- Introduction
- Methods
  - Experimental Setup
  - Analyzed Models
- Result
  - Constituency Label Prediction
  - Additional Task: Dependency Arc Prediction
- Conclusion

Introduction

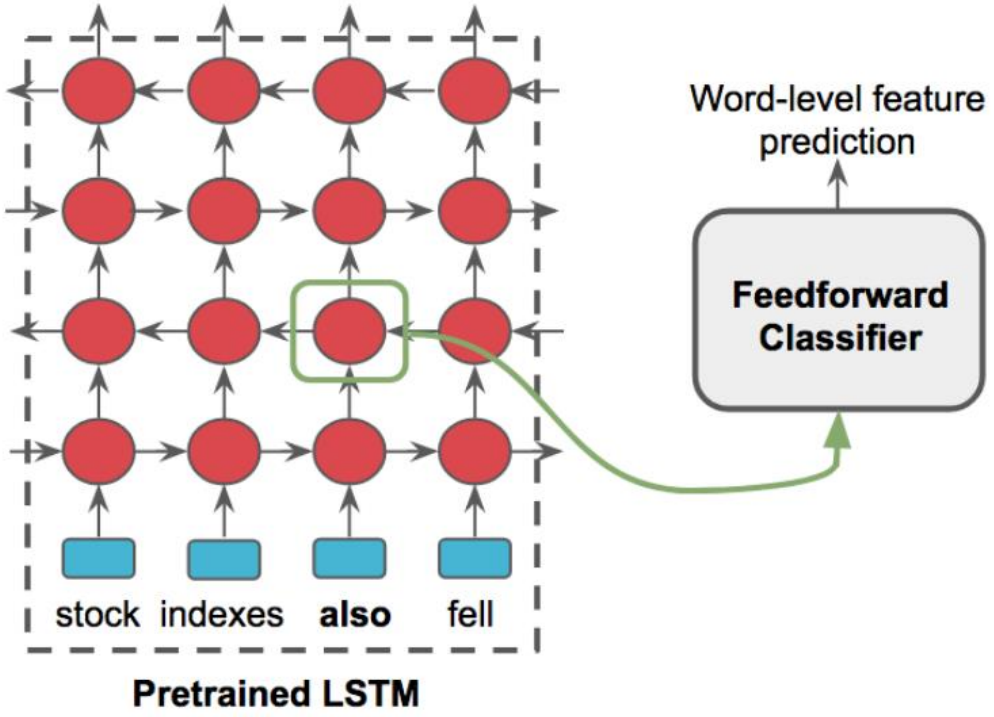
- Recurrent Neural Networks (RNNs) have surpassed NLP models using explicit syntactic features (e.g. POS, dependencies)
- Nevertheless, some models benefit from syntactic features or supervision, e.g. in Neural Machine Translation or Question Answering ([3], [4])

→What do RNNs learn to represent internally?

- We will see that no explicit syntactic supervision is needed!

Experimental Setup

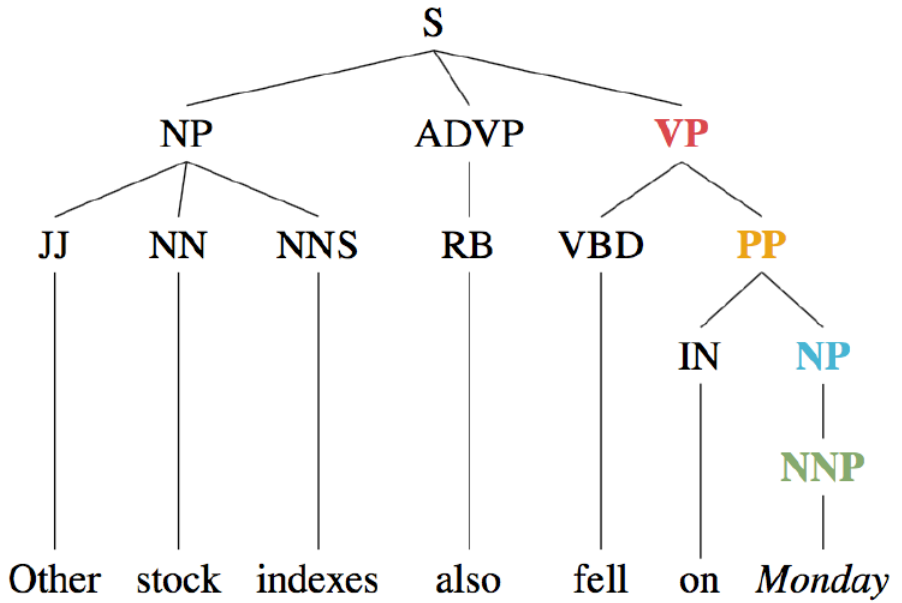
- 4 multi-layered RNNS under investigation



- $x_i^l = vector\ representation\ \ of\ \ each\ word\ i\ at\ layer\ l$

Experimental Setup

- Predict POS, parent, grandparent and great-grandparent constituent label for each x_i^l



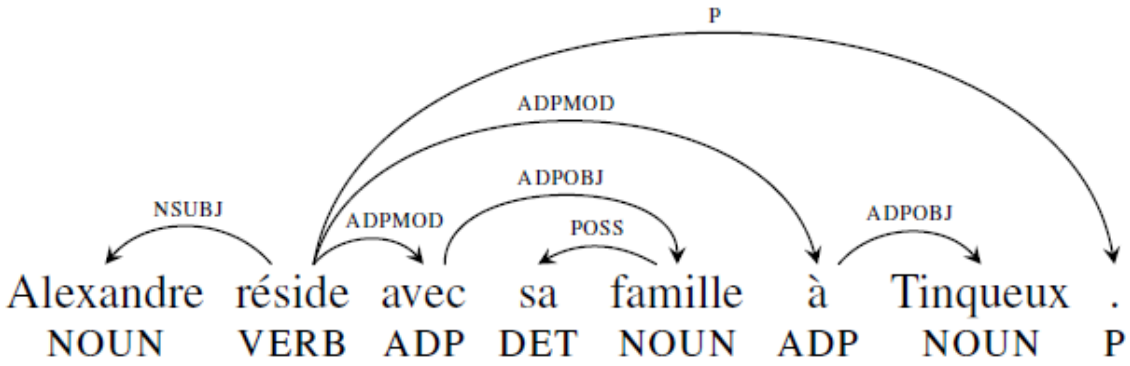
Experimental Setup

- Feed-forward neural network with single 300-dimensional hidden layer
- Activation function: $y_i^l = SoftMax(W_2 ReLu(W_1\ x_i^l))$
- One classifier per layer and prediction task
- trained on dev set of CoNLL-2012 and evaluated on test set
- Baseline: majority class prediction (which outperforms the classifier trained on pretrained GloVe embeddings)

Analyzed Models

- Four different forms of supervision:
  - Dependency Parsing, Semantic Role Labeling, Machine Translation and Language Modeling
- They changed some parameters (and do not state which), but use mainly standard ones

Dependency Parsing



One sample sentence from French UD [1]

- **Stanford Dependency Parser**
- 4-layer bidirectional LSTM
- 400-dimensional hidden units
- Trained on English Universal Dependency Web Treebank
- UAS 91.5 / LAS 82.18

- POS Tags as features → should contain a high amount of syntactic information

Semantic Role Labeling

...the company to offer a 15% stake to the public.

Arg0:	the company
Rel:	offer
Arg1:	a 15% stake
Arg2−to:	the public

One sample relationship in CoNLL-2012 [2]

- Pretrained DeepSRL model
- 8-layer alternating directions LSTMs with highways
- 300-dimensional hidden units
- trained on CoNLL-2012 training set
- Concatenation of each forward/backward-layer pair

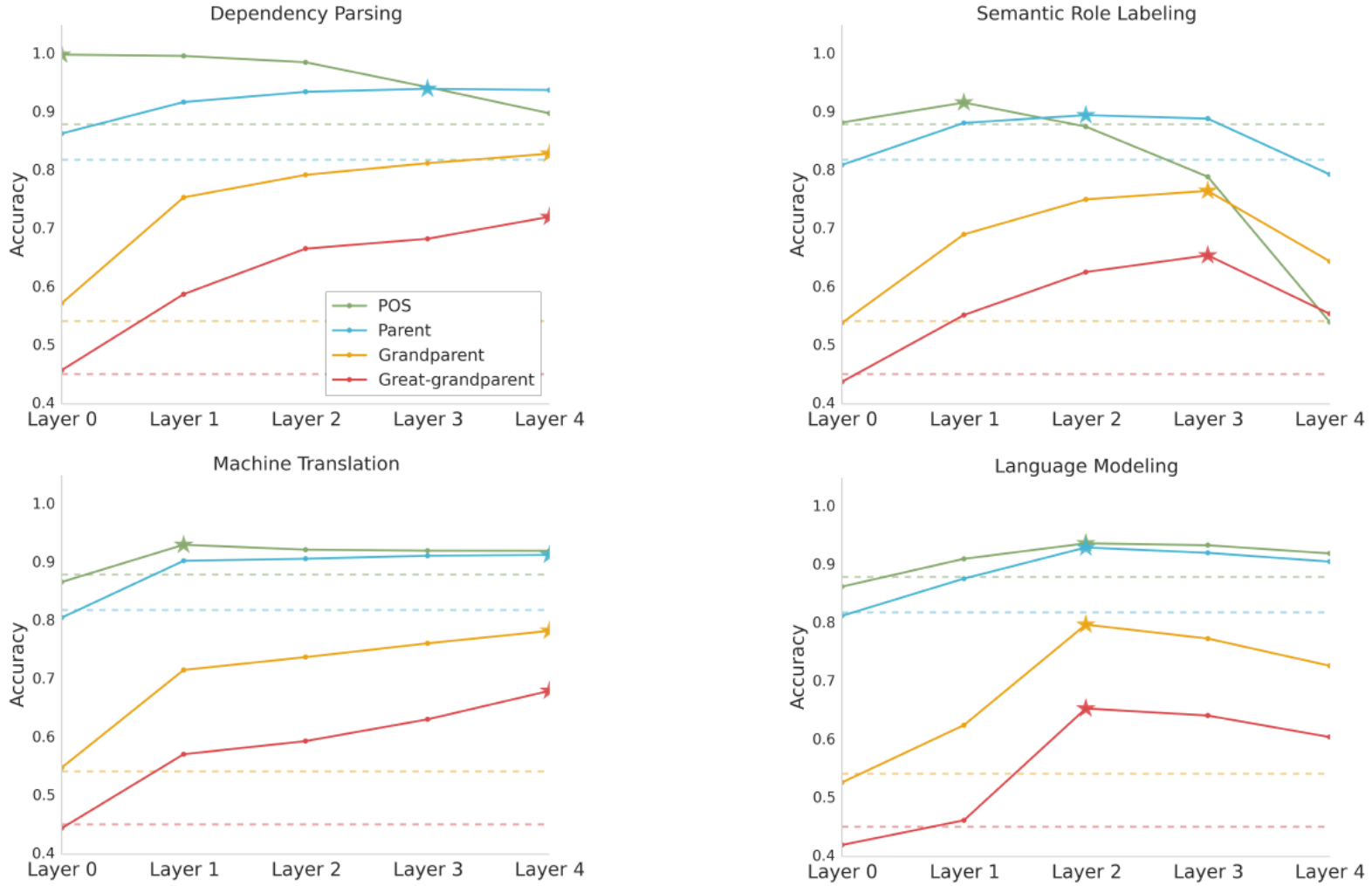
Machine Translation

- Using OpenNMT model
- 4-layer bidirectional LSTM
- 500-dimensional hidden units
- Trained on WMT-14 English-German dataset
- BLEU score of 21.37 (Transformer has 35 points on same dataset)

Language Modeling

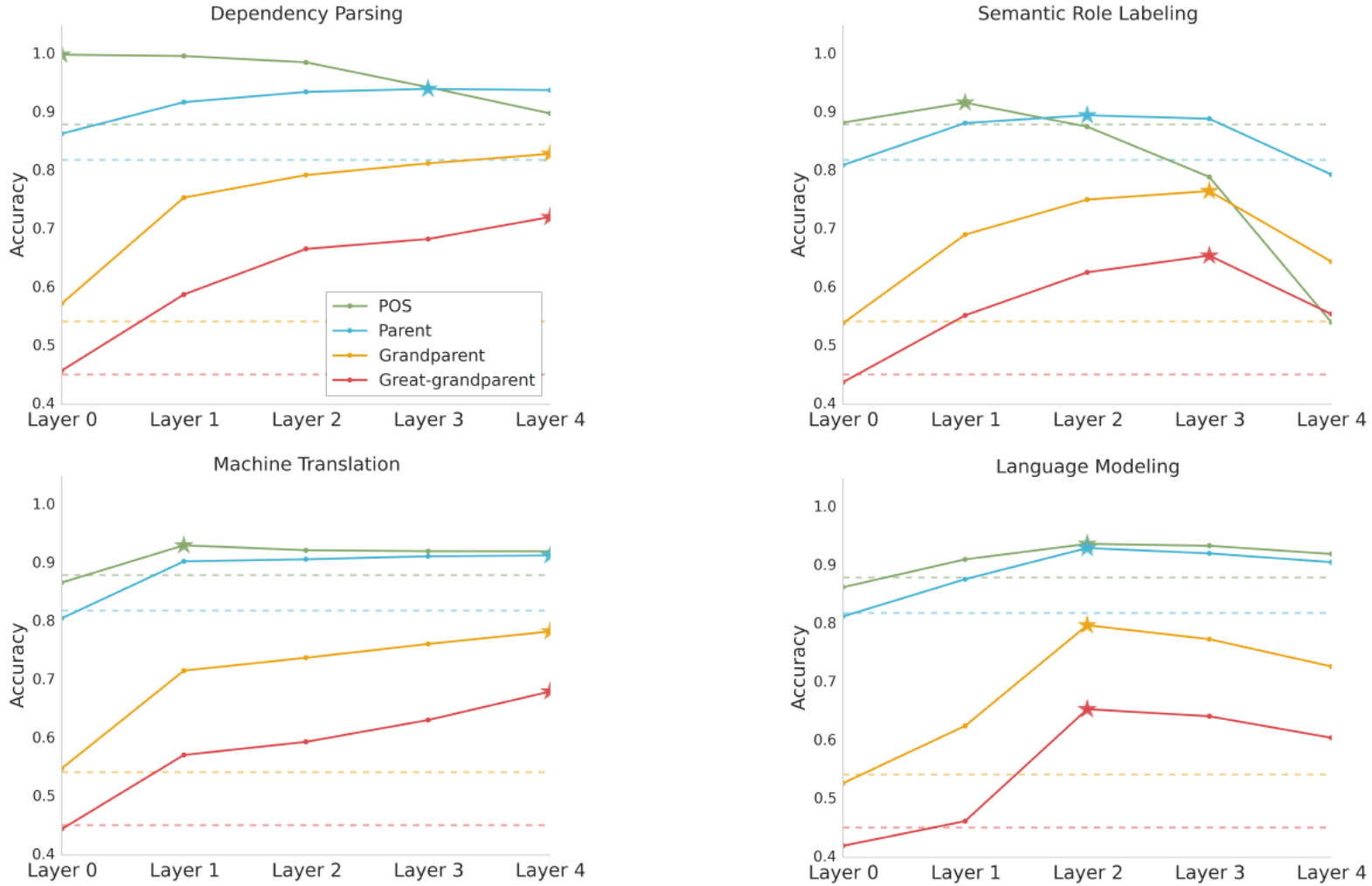
- one forward and one backward 4-layer LSTM with highways
- 1000-dimensional hidden units
- variational dropout and tied input-output embeddings
- Trained on CoNLL-2012 training set
- Concatenation of respective forward and backward layer representation

Result: Constituency Label Prediction



Results of syntax experiments. The best performing layer for each experiment is annotated with a star, and the per-word majority baseline for each task is shown with a dashed line.

- RNNs can induce syntax



Results of syntax experiments. The best performing layer for each experiment is annotated with a star, and the per-word majority baseline for each task is shown with a dashed line.

- Deeper layers reflect higher-level syntax

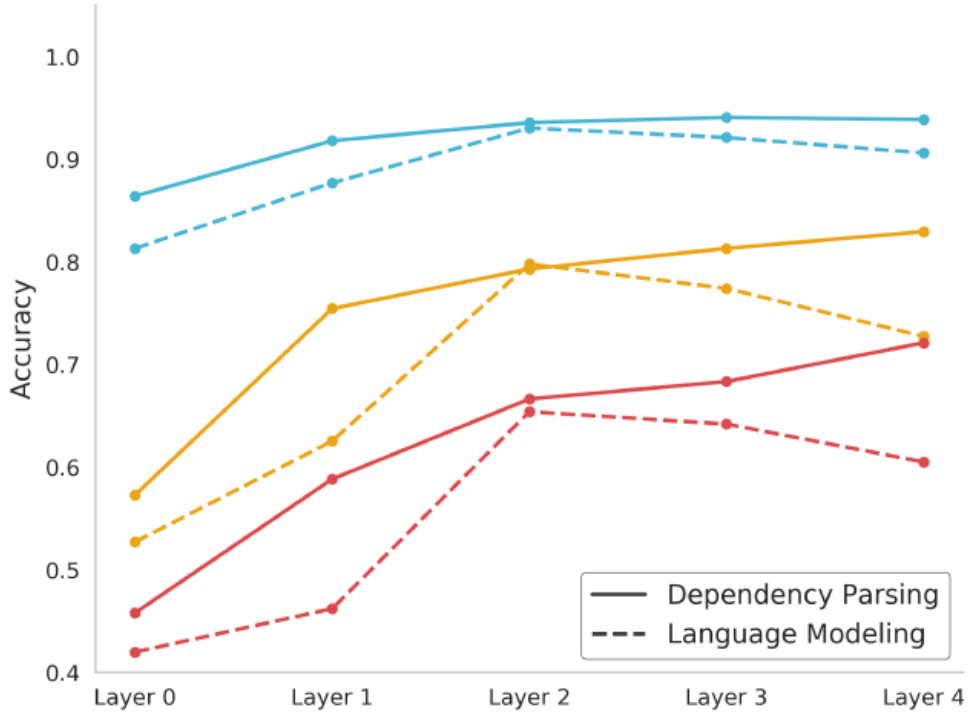
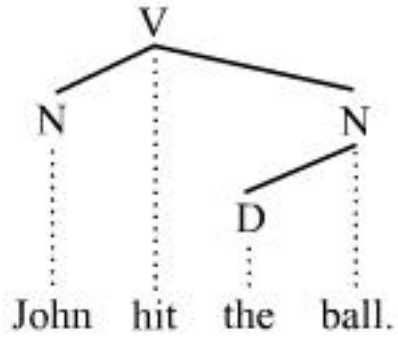


Figure 3: Comparison between the LM and dependency parser on the parent (blue), grandparent (yellow), and great-grandparent (red) constituent prediction tasks.

- Linzen et al. : classifier trained on subject-verb agreement using representations from a LM performs worse than baseline ([5])
- LM model does learn syntactic features

Additional Task: Dependency Arc Prediction

- Do two words share a dependency arc?



Dependency-based parse tree

- Same models as before
- $input = [w_p; w_c; w_p o\ w_C]$
- UD Web Treebank, trained on dev set and tested on test set
- For each word in UD dataset:
  - word + parent
  - word + random word

Results: Dependency Arc Prediction

Source Model	GloVe	L0	L1	L2	L3	L4
DP	0.50	0.68	0.77	0.81	0.88	**0.95**
SRL	0.50	0.58	0.69	0.76	**0.79**	0.74
MT	0.50	0.61	**0.73**	0.63	0.63	0.63
LM	0.50	0.62	0.74	0.78	**0.80**	0.73

Table 2: Results of the dependency arc prediction task. L0–L4 denote the different layers of the model. DP refers to the RNN trained with dependency parsing supervision.

Conclusion

- Deep RNNs can learn syntax without explicit supervision

- Deeper layers contain higher-level syntactic information

  → soft hierarchy over syntax established

- Results indicate the power of Transfer Learning

→Use representations of LSTMs as input to other RNNs, CNNs or networks with attention

References

1) McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., ... & Bedini, C. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 92-97).

2) Kingsbury, P., & Palmer, M. (2002, May). From TreeBank to PropBank. In *LREC* (pp. 1989-1993).

3) Roee Aharoni and Yoav Goldberg. 2017. Towards string-to-tree neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, pages 132–140. https://doi.org/10.18653/v1/P17-2021.

4) Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

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