Syntax in neural language models: a case study

Dieuwke Hupkes

Institute for Logic, Language and Computation
University of Amsterdam

Universiteit Utrecht

November 15, 2019

1. They should have some desired properties w.r.t what you want to understand;

Neural networks as explanatory models

Prerequisites

1. They should have some desired properties w.r.t what you want to understand;
2. They should be adequate models of the phenomenon that you are interested in;

Neural networks as explanatory models

Prerequisites

1. They should have some desired properties w.r.t what you want to understand;
2. They should be adequate models of the phenomenon that you are interested in;
3. You should be able to obtain insight into *how* they model this phenomenon.

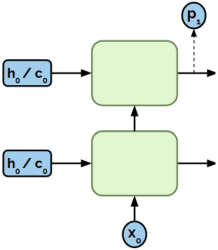
Artificial languages

- The compositionality of neural networks: integrating symbolism and connectionism *(Hupkes et al. 2019b)*
- Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure *(Hupkes, Veldhoen, and Zuidema 2018)*
- Learning compositionally through attentive guidance *(Hupkes et al. 2019a)*
- Diagnostic classification and symbolic guidance to understand and improve recurrent neural networks *(Hupkes and Zuidema 2017)*

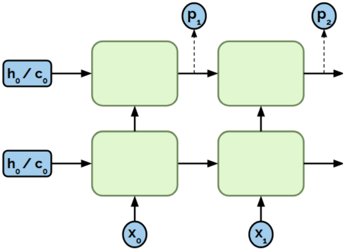
Language modelling

- Under the hood: using diagnostic classifiers to investigate and improve how language models track agreement information *(Giulianelli, Harding, Mohnert, Hupkes and Zuidema, 2018)*
- The emergence of number and syntax units in LSTM language models *(Lakretz, Kruszewski, Desbordes, Hupkes, Dehaene and Baroni, 2019)*
- Analysing neural language models: contextual decomposition reveals default reasoning in number and gender assignment *(Jumelet, Zuidema and Hupkes, 2019)*

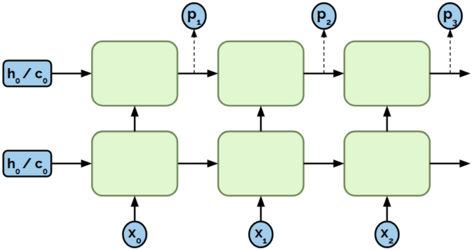
Language modelling



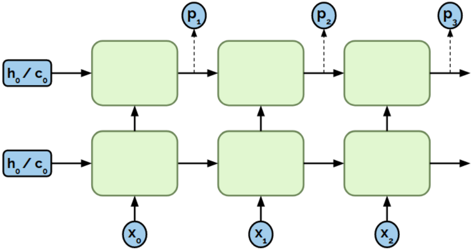
Language modelling



Language modelling



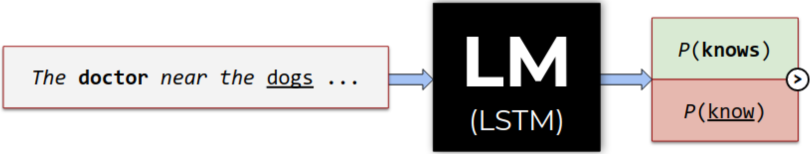
Language modelling



- 2-layer LSTM model
- Trained data: 90M Wikipedia tokens

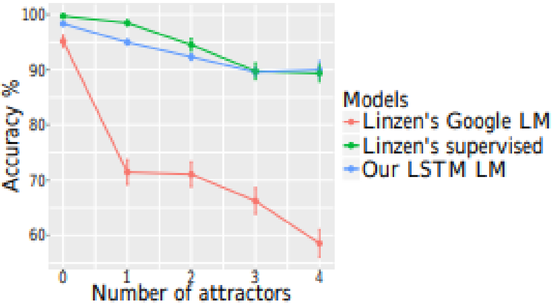
- Captures non-trivial aspects of syntactic structure!

Subject-verb agreement



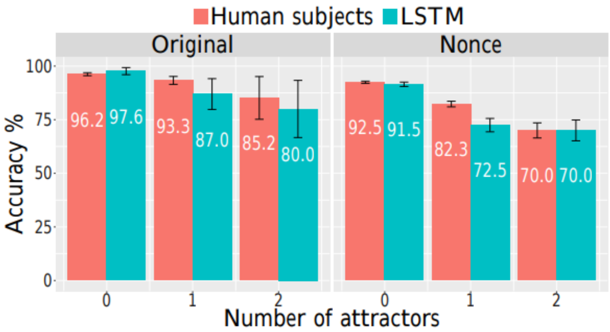
(Linzen, Dupoux, and Goldberg 2016)

Results



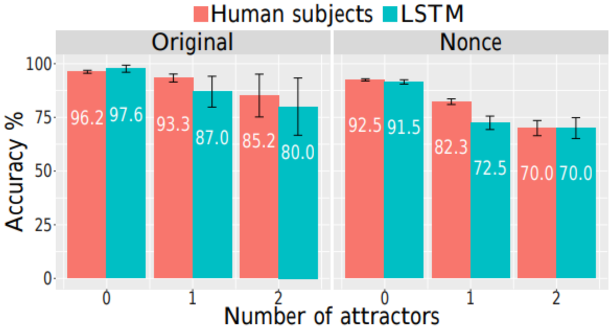
(Gulordava et al. 2018)

Original and nonsensical sentences



(Gulordava et al. 2018)

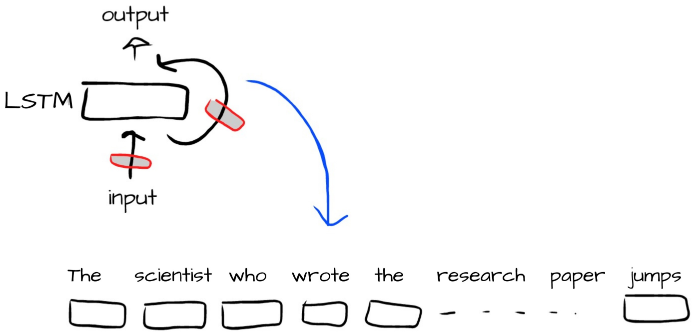
Original and nonsensical sentences



- *How* do they do this?

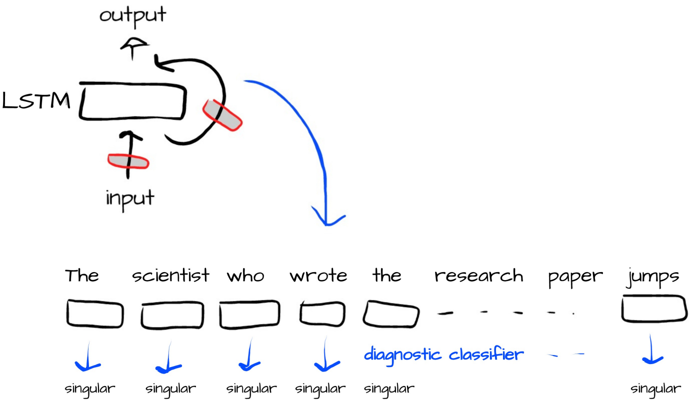
Diagnostic Classification

Diagnostic Classification



(Hupkes, Veldhoen, and Zuidema 2018; Veldhoen, Hupkes, and Zuidema 2016)

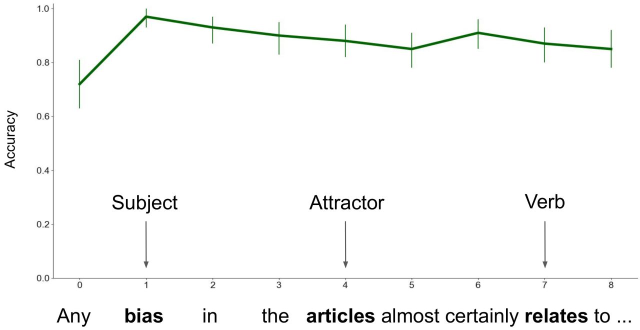
Diagnostic Classification



(Hupkes, Veldhoen, and Zuidema 2018; Veldhoen, Hupkes, and Zuidema 2016)

Diagnostic Classification

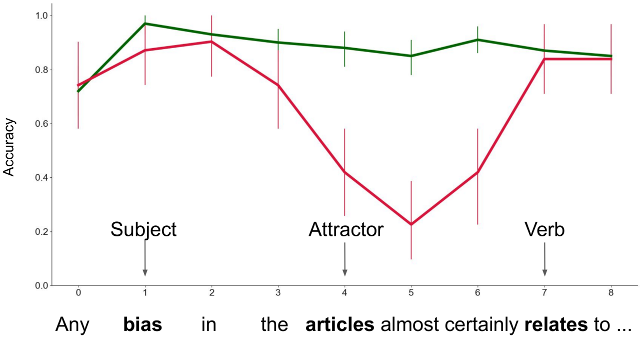
Sentences with correct predictions, h



(Giulianelli, Harding, Mohnert, Hupkes and Zuidema)

Diagnostic Classification

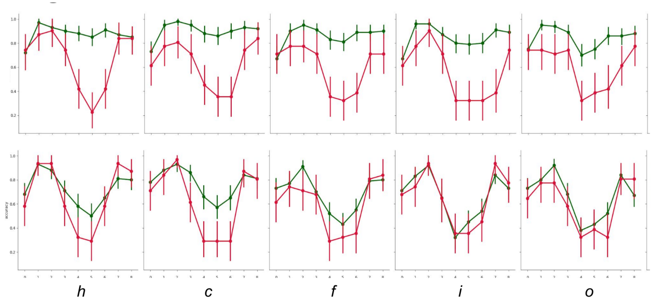
All sentences, h



(Giulianelli et al. 2018)

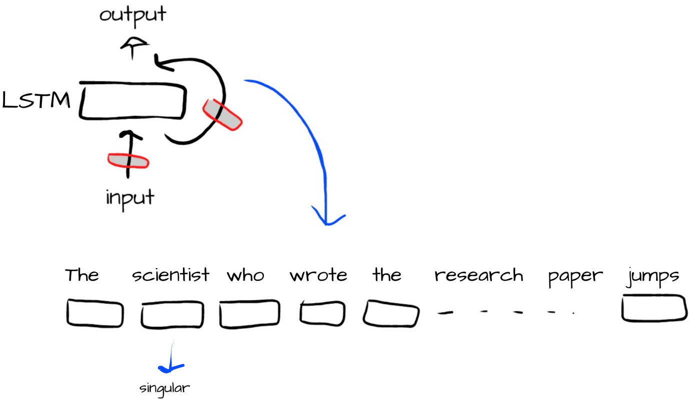
Diagnostic Classification

All sentences, all components



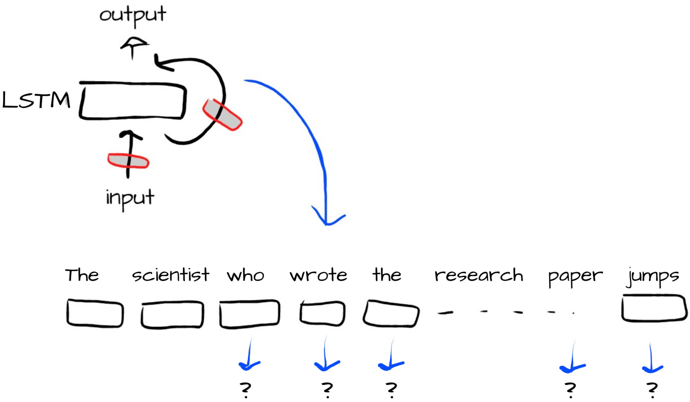
(Giulianelli et al. 2018)

Temporal Generalisation



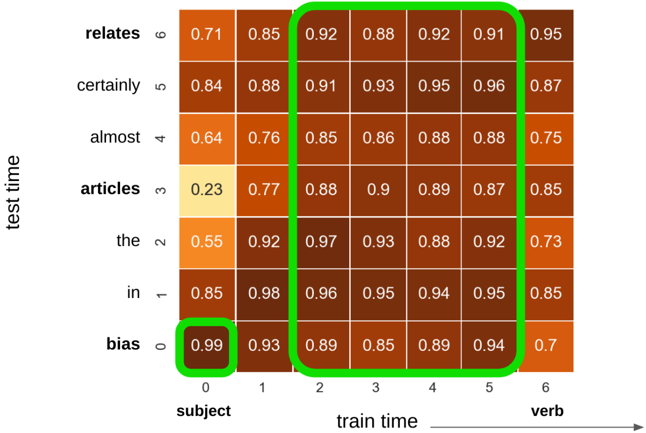
(Giulianelli et al. 2018)

Temporal Generalisation



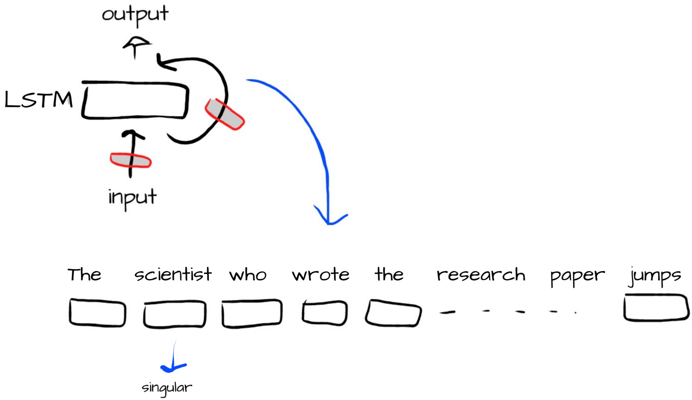
(Giulianelli et al. 2018)

Temporal generalisation matrix



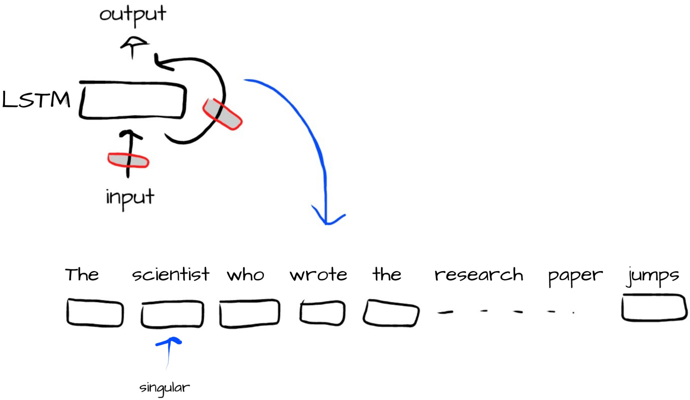
(Giulianelli et al. 2018)

Diagnostic interventions



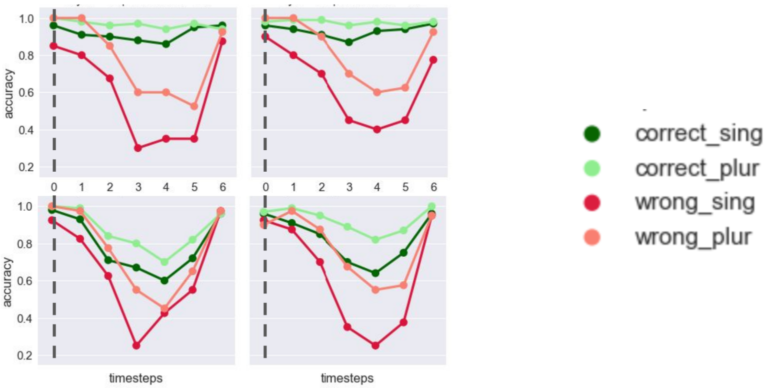
(Giulianelli et al. 2018)

Diagnostic interventions



(Giulianelli et al. 2018)

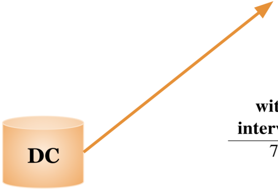
Diagnostic Interventions



(Giulianelli et al. 2018)

Diagnostic interventions, results

	An	official	estimate	issued	in	2003	suggests	suggest
Original		-11.05	-8.426	-8.472	-1.243	-3.951	-5.753	**-5.6979**
Intervention		-11.05	-8.426	-8.472	-1.268	-3.97	**-5.691**	-6.4361



without intervention	**with intervention**
78.0	85.4

* Overall differences in sentence perplexities are statistically insignificant

(Giulianelli et al. 2018)

Conclusions

With *Diagnostic Classification* we can discover if, when and where information is represented in a recurrent neural network:

Conclusions

With *Diagnostic Classification* we can discover if, when and where information is represented in a recurrent neural network:

- Number information is stored mostly in the hidden and cell states of the LSTM language model;

Conclusions

With *Diagnostic Classification* we can discover if, when and where information is represented in a recurrent neural network:

- Number information is stored mostly in the hidden and cell states of the LSTM language model;
- The model maintains a *deep* and *surface* representation of number;

Conclusions

With *Diagnostic Classification* we can discover if, when and where information is represented in a recurrent neural network:

- Number information is stored mostly in the hidden and cell states of the LSTM language model;
- The model maintains a *deep* and *surface* representation of number;
- The model is indeed distracted by the attractor, but for wrong trials, the encoding already goes wrong *before* the attractor;

Conclusions

With *Diagnostic Classification* we can discover if, when and where information is represented in a recurrent neural network:

- Number information is stored mostly in the hidden and cell states of the LSTM language model;
- The model maintains a *deep* and *surface* representation of number;
- The model is indeed distracted by the attractor, but for wrong trials, the encoding already goes wrong *before* the attractor;
- We can influence the behaviour of the model by *inverting* the diagnostic classifiers.

Ablation Studies

Templates for number-agreement tasks

Simple
Adv
2Adv
CoAdv
NamePP
NounPP
NounPPAdv

(Lakretz, Kruszewski, Desbordes, Hupkes, Dehaene and Baroni)

Templates for number-agreement tasks

Simple	the **boy greets** the guy
Adv	
2Adv	
CoAdv	
NamePP	
NounPP	
NounPPAdv	

(Lakretz, Kruszewski, Desbordes, Hupkes, Dehaene and Baroni)

Templates for number-agreement tasks

Simple	the **boy** **greets** the guy
Adv	the **boy** probably **greets** the guy
2Adv	
CoAdv	
NamePP	
NounPP	
NounPPAdv	

(Lakretz, Kruszewski, Desbordes, Hupkes, Dehaene and Baroni)

Templates for number-agreement tasks

Simple	the **boy greets** the guy
Adv	the **boy** probably **greets** the guy
2Adv	the **boy** most probably **greets** the guy
CoAdv	
NamePP	
NounPP	
NounPPAdv	

(Lakretz, Kruszewski, Desbordes, Hupkes, Dehaene and Baroni)

Templates for number-agreement tasks

Simple	the **boy** **greets** the guy
Adv	the **boy** probably **greets** the guy
2Adv	the **boy** most probably **greets** the guy
CoAdv	the **boy** openly and deliberately **greets** the guy
NamePP	
NounPP	
NounPPAdv	

(Lakretz, Kruszewski, Desbordes, Hupkes, Dehaene and Baroni)

Templates for number-agreement tasks

Simple	the **boy** **greets** the guy
Adv	the **boy** probably **greets** the guy
2Adv	the **boy** most probably **greets** the guy
CoAdv	the **boy** openly and deliberately **greets** the guy
NamePP	the **boy** near Pat **greets** the guy
NounPP	
NounPPAdv	

(Lakretz, Kruszewski, Desbordes, Hupkes, Dehaene and Baroni)

Templates for number-agreement tasks

Simple	the **boy** **greets** the guy
Adv	the **boy** probably **greets** the guy
2Adv	the **boy** most probably **greets** the guy
CoAdv	the **boy** openly and deliberately **greets** the guy
NamePP	the **boy** near Pat **greets** the guy
NounPP	the **boy** near the car **greets** the guy
NounPPAdv	

(Lakretz, Kruszewski, Desbordes, Hupkes, Dehaene and Baroni)

Templates for number-agreement tasks

Simple	the **boy** **greets** the guy
Adv	the **boy** probably **greets** the guy
2Adv	the **boy** most probably **greets** the guy
CoAdv	the **boy** openly and deliberately **greets** the guy
NamePP	the **boy** near Pat **greets** the guy
NounPP	the **boy** near the car **greets** the guy
NounPPAdv	the **boy** near the car kindly **greets** the guy

(Lakretz, Kruszewski, Desbordes, Hupkes, Dehaene and Baroni)

Ablation Results

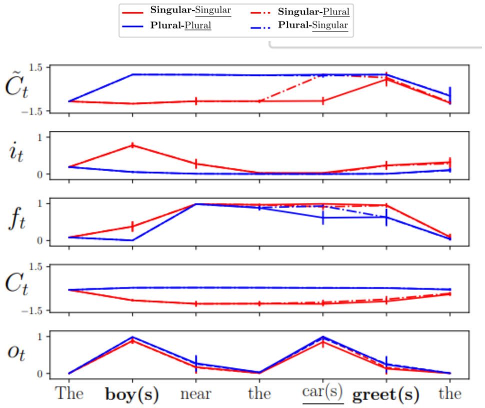
NA task	Condition	Full Model
Simple	S	100
Adv	S	100
2Adv	S	99.9
CoAdv	S	98.7
namePP	SS	99.3
nounPP	SS	99.2
nounPP	SP	87.2
nounPPAdv	SS	99.5
nounPPAdv	SP	91.2
Simple	P	100
Adv	P	99.6
2Adv	P	99.3
CoAdv	P	99.3
namePP	PS	68.9
nounPP	PS	92.0
nounPP	PP	99.0
nounPPAdv	PS	99.2
nounPPAdv	PP	99.8

Ablation Results

NA task	Condition	Full Model	Ablated 776	Ablated 988
Simple	S	100	-	-
Adv	S	100	-	-
2Adv	S	99.9	-	-
CoAdv	S	98.7	-	82
namePP	SS	99.3	-	-
nounPP	SS	99.2	-	-
nounPP	SP	87.2	-	54.2
nounPPAdv	SS	99.5	-	-
nounPPAdv	SP	91.2	-	54.0
Simple	P	100	-	-
Adv	P	99.6	-	-
2Adv	P	99.3	-	-
CoAdv	P	99.3	79.2	-
namePP	PS	68.9	39.9	-
nounPP	PS	92.0	48.0	-
nounPP	PP	99.0	78.3	-
nounPPAdv	PS	99.2	63.7	-
nounPPAdv	PP	99.8	-	-

Singular unit behaviour

$$c_t = f_t \circ c_{t-1} + i_t \circ \widetilde{c}_t$$
$$h_t = o_t \circ \tanh(c_t)$$



(a) 988 (singular)

(Lakretz et al. 2019)

Diagnostic Classification 2

- Short distance relations?

Diagnostic Classification 2

- Short distance relations?
  - → Diagnostic classifiers to predict *number* information
  - → Ablation to confirm the role of short range units

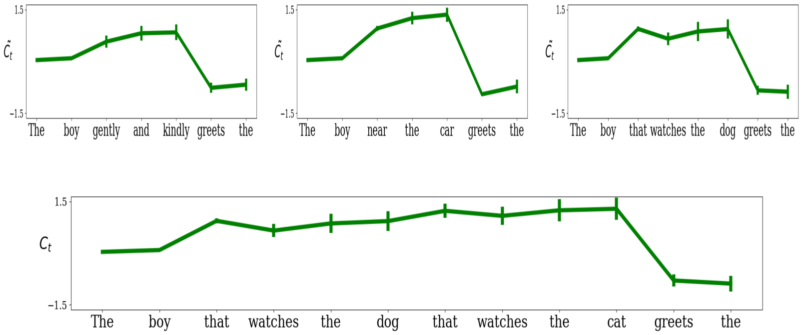
Diagnostic Classification 2

- Short distance relations?
  - $\rightarrow$ Diagnostic classifiers to predict *number* information
  - $\rightarrow$ Ablation to confirm the role of short range units
- The syntactic structure?

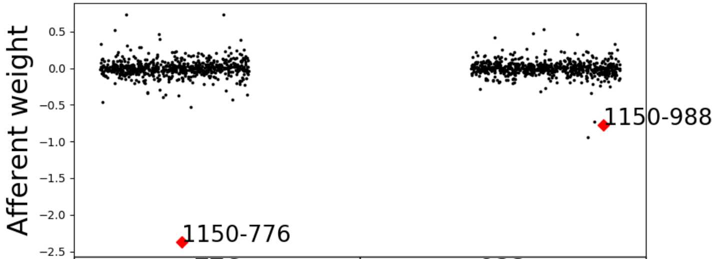
Diagnostic Classification 2

- Short distance relations?
  - $\rightarrow$ Diagnostic classifiers to predict *number* information
  - $\rightarrow$ Ablation to confirm the role of short range units
- The syntactic structure?
  - $\rightarrow$ Diagnostic classifiers to predict *syntactic depth*
  - $\rightarrow$ Ablation to confirm the role of the syntax units

Syntax unit 1150, cell activity



Syntax unit 1150, outgoing weights



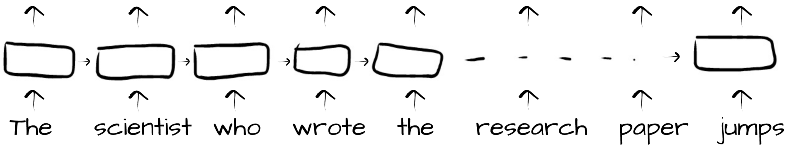
	776	988
776		0.53
988	0.09	
1150	-2.37	-0.77

Conclusions

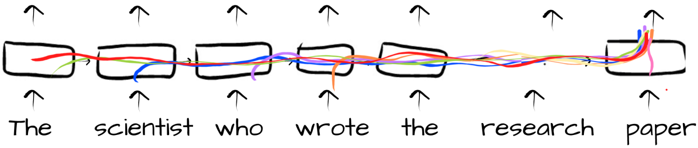
- Using **ablation**, we found that long distance number is encoded locally, in two units;
    - One *singular* unit
    - One *plural* unit
- Using **diagnostic classifiers and ablation**, we found that short distance number is encoded in a distributed fashion;
- Using **diagnostic classification**, we found a number of syntax units, one of which highly interpretable.

Generalised Contextual Decomposition

Contextual Decomposition



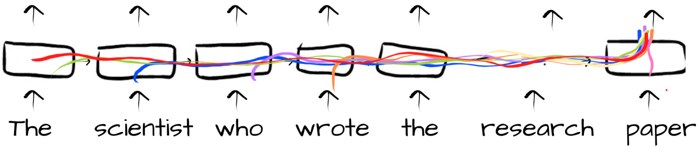
Contextual Decomposition



- Keep track of interactions

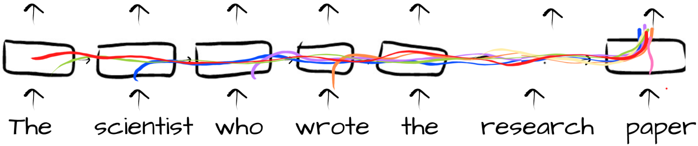
(Murdoch, Liu, and Yu 2018)

Contextual Decomposition



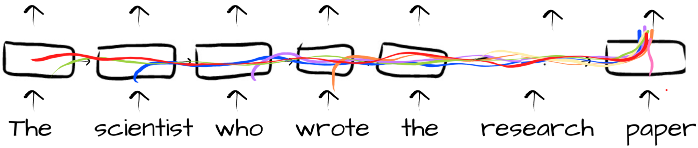
- Keep track of interactions
  - Linear sums: 3 * **2** + 1 * **4**

Contextual Decomposition



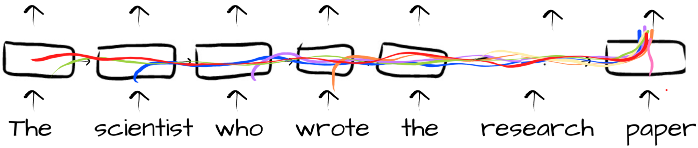
- Keep track of interactions
  - Linear sums: $3 * \mathbf{2} + 1 * \mathbf{4}$
  - Non-linearities: $\textsc{tanh}(\mathbf{10} + \mathbf{20})$

Contextual Decomposition



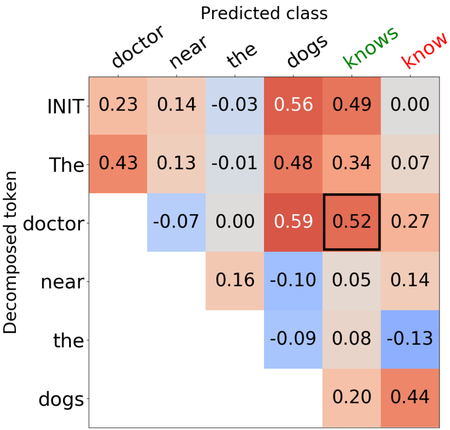
- Keep track of interactions
  - Linear sums: $3 * \mathbf{2} + 1 * \mathbf{4}$
  - Non-linearities: $\textsc{tanh}(\mathbf{10} + \mathbf{20}) \rightarrow$ **Shapley decompositions**
  - Multiplications: $\mathbf{5} * \mathbf{2}$

Contextual Decomposition



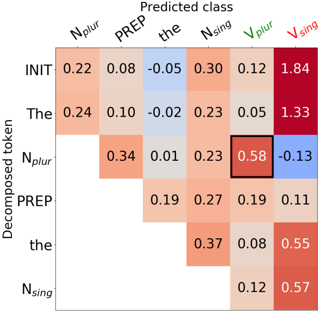
- Keep track of interactions
  - Linear sums: $3 * \mathbf{2} + 1 * \mathbf{4}$
  - Non-linearities: $\textsc{tanh}(\mathbf{10} + \mathbf{20})$
  - Multiplications: $\mathbf{5} * \mathbf{2}$
- Which interactions?

Information flow "attention" plots



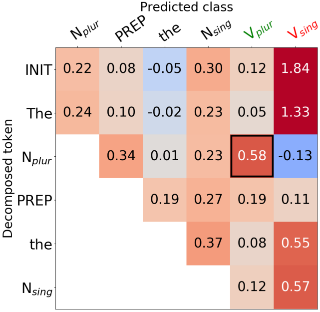
(Jumelet, Hupkes, and Zuidema 2019)

Singular versus plural

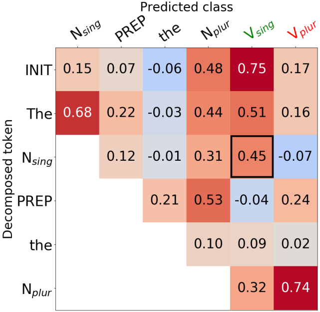


NounPP – PS

Singular versus plural



NounPP – PS



NounPP – SP

Pruning information

Task	Condition	FULL	GCD IN
Simple	S	100	73.3
Simple	P	100	100
nounPP	SS	99.2	93.0
nounPP	SP	87.2	90.3
nounPP	PS	92.0	100
nounPP	PP	99.0	100
namePP	SS	99.3	97.7
namePP	PS	68.9	98.3

- FULL: full model accuracy
- IN: information from the subject,

Pruning information

Task	Condition	FULL	GCD IN	GCD INTERCEPT*
Simple	S	100	73.3	97.3
Simple	P	100	100	32.7
nounPP	SS	99.2	93.0	99.8
nounPP	SP	87.2	90.3	98.8
nounPP	PS	92.0	100	0.0
nounPP	PP	99.0	100	7.0
namePP	SS	99.3	97.7	99.4
namePP	PS	68.9	98.3	1.3

- FULL: full model accuracy
- IN: information from the subject,
- INTERCEPT*: only intercept interactions

Pruning information

			GCD		
Task	**Condition**	FULL	IN	INTERCEPT*	¬INTERCEPT
Simple	S	100	73.3	97.3	69.7
Simple	P	100	100	32.7	100
nounPP	SS	99.2	93.0	99.8	72.7
nounPP	SP	87.2	90.3	98.8	60.5
nounPP	PS	92.0	100	0.0	100
nounPP	PP	99.0	100	7.0	99.8
namePP	SS	99.3	97.7	99.4	76.2
namePP	PS	68.9	98.3	1.3	99.9

- FULL: full model accuracy
- IN: information from the subject,
- INTERCEPT*: only intercept interactions
- ¬INTERCEPT: no intercept interactions

Conclusions

We can use contextual decomposition to track the information flow in recurrent neural networks:

- Plural verbs have a much stronger causal relationship to their plural subject than singular verbs to their singular subject.
- By considering different types of interactions, we find that to predict singular verbs, the model relies heavily on its intercepts
- GCD can also be used in other kinds of scenario's, where behavioural accuracy tests are not possible (anaphora resolution, negative polarity items)!

What's next?

What's next?

- Other linguistic questions

What's next?

- Other linguistic questions
  - Negative polarity items (Jumelet and Hupkes 2018; Marvin and Linzen 2018)
  - Filler-gap dependencies (Wilcox et al. 2018, 2019)
  - Reflexive anaphora (Futrell et al. 2019; Jumelet, Hupkes, and Zuidema 2019; Marvin and Linzen 2018)
  - Garden path sentences (Futrell et al. 2019; Van Schijndel and Linzen 2018; Wilcox et al. 2019)
  - Syntactic priming (Prasad, Schijndel, and Linzen 2019; Van Schijndel and Linzen 2018)
  - And many more...

- Other "model" questions

What's next?

- Other linguistic questions
- Other "model" questions
  - Do structural biases help? (Futrell et al. 2018; Wilcox et al. 2019)
  - What is the impact of quantity and quality of training data (Schijndel, Mueller, and Linzen 2019)?

What's next?

- Other linguistic questions
- Other "model" questions
- The ultimate question

What's next?

- Other linguistic questions
- Other "model" questions
- The ultimate question
  - How does this help us to better understand human language processing?

What's next?

- Other linguistic questions
- Other "model" questions
- The ultimate question
  - How does this help us to better understand human language processing?

I'm looking forward to figuring those things out!

Thanks to my collaborators

Willem Zuidema

Marco Baroni

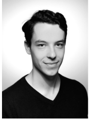
Jaap Jumelet

Germán Kruszewski

Yair Lakretz

Sara Veldhoen

Mario Giulianelli



Florian Mohnert

Jack Harding

Thank you

Thank you for your attention!

ILLC   UvA

dieuwkehupkes@gmail.com
https://dieuwkehupkes.nl
https://www.instagram.com/duo_polenotti/

Kyunghyun Cho et al. "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches". In: *SSST@EMNLP*. 2014, pp. 103–111.

Junyoung Chung et al. "Gated feedback recurrent neural networks". In: *ICML*. 2015, pp. 2067–2075.

Jeffrey L Elman. "Finding structure in time". In: *Cognitive science* 14.2 (1990), pp. 179–211.

Richard Futrell et al. "RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency". In: *arXiv preprint arXiv:1809.01329* (2018).

Richard Futrell et al. "Neural language models as psycholinguistic subjects: Representations of syntactic state". In: *NAACL*. Association for Computational Linguistics, 2019, pp. 32–42.

References II

Mario Giulianelli et al. "Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2018, pp. 240–248.

Kristina Gulordava et al. "Colorless Green Recurrent Networks Dream Hierarchically". In: *Proceedings of NAACL*. Vol. 1. 2018, pp. 1195–1205.

Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

References III

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. "Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure". In: *Journal of Artificial Intelligence Research* 61 (2018), pp. 907–926.

Dieuwke Hupkes and Willem Zuidema. "Diagnostic classification and symbolic guidance to understand and improve recurrent neural networks". In: *Proceedings Workshop on Interpreting, Explaining and Visualizing Deep Learning, NIPS2017*. 2017.

Dieuwke Hupkes et al. "Learning compositionally through attentive guidance". In: *Proceedings of Cicling*. 2019.

References IV

Dieuwke Hupkes et al. "The compositionality of neural networks: integrating symbolism and connectionism". In: *CoRR* abs/1908.08351 (2019).

Jaap Jumelet and Dieuwke Hupkes. "Do Language Models Understand Anything? On the Ability of LSTMs to Understand Negative Polarity Items". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2018, pp. 222–231.

Jaap Jumelet, Dieuwke Hupkes, and Willem Zuidema. "Analysing Neural Language Models: Contextual Decomposition Reveals Default Reasoning in Number and Gender Assignment". In: *CoNLL*. 2019.

References V

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. “Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies”. In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 521–535. ISSN: 2307-387X.

Yair Lakretz et al. “The emergence of number and syntax units in LSTM language models”. In: *arXiv preprint arXiv:1903.07435* (2019).

Rebecca Marvin and Tal Linzen. “Targeted Syntactic Evaluation of Language Models”. In: *EMNLP*. 2018, pp. 1192–1202.

W. James Murdoch, Peter J. Liu, and Bin Yu. “Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs”. In: *ICLR*. 2018.

References VI

Grusha Prasad, Marten van Schijndel, and Tal Linzen. "Using Priming to Uncover the Organization of Syntactic Representations in Neural Language Models". In: *CoNLL*. 2019.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. "Quantity doesn't buy quality syntax with neural language models". In: *CoRR* abs/1909.00111 (2019).

Sara Veldhoen, Dieuwke Hupkes, and Willem Zuidema. "Diagnostic Classifiers: Revealing how Neural Networks Process Hierarchical Structure". In: *Pre-Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (CoCo @ NIPS 2016)*. 2016.

References VII

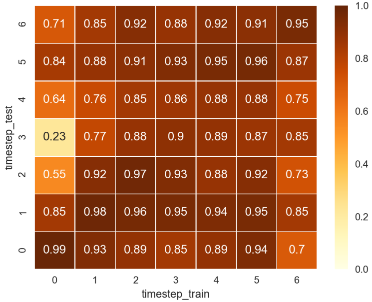
Marten Van Schijndel and Tal Linzen. "Modeling garden path effects without explicit hierarchical syntax.". In: *CogSci*. 2018.

Ethan Wilcox et al. "What do RNN language models learn about filler–gap dependencies?" In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2018, pp. 211–221.

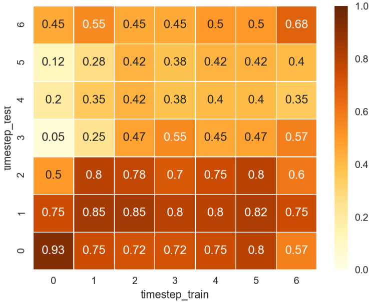
Ethan Wilcox et al. "Structural Supervision Improves Learning of Non-Local Grammatical Dependencies". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 3302–3312.

Appendices

Temporal Generalisation

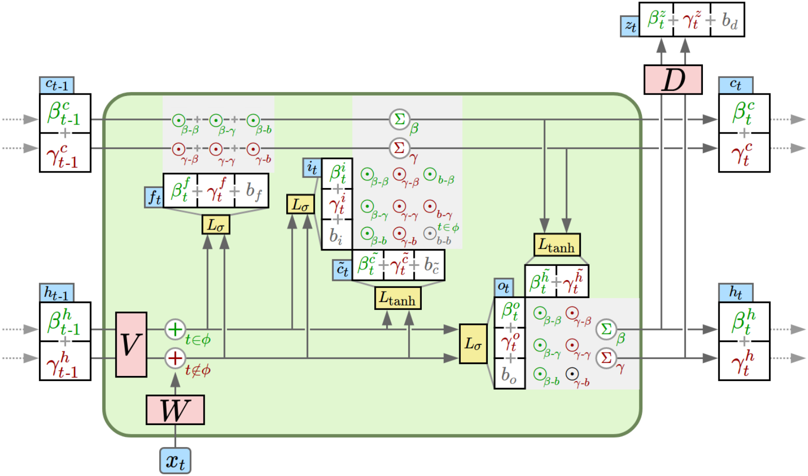


Correct trials



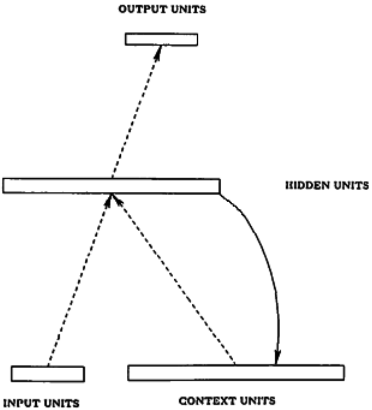
Wrong trials

Generalised Contextual Decomposition



Simple Recurrent Network

$$\mathbf{h}_t = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{\mathbf{t-1}} + \mathbf{b})$$

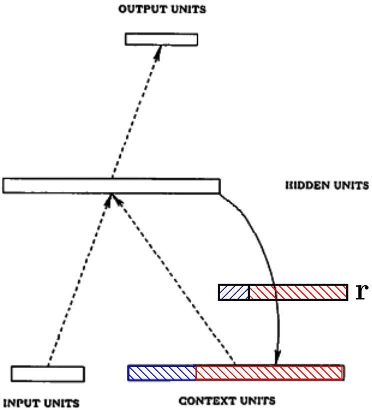


(Elman 1990)

Gated recurrent neural networks

$$\mathbf{h}_t = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{\mathbf{t-1}} + \mathbf{b})$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1} + \mathbf{b}_r)$$

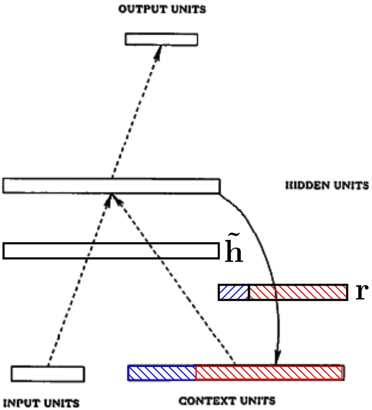


(Cho et al. 2014; Chung et al. 2015)

Gated recurrent neural networks

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r} \odot \mathbf{h}_{t-1}) + \mathbf{b})$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1} + \mathbf{b}_r)$$



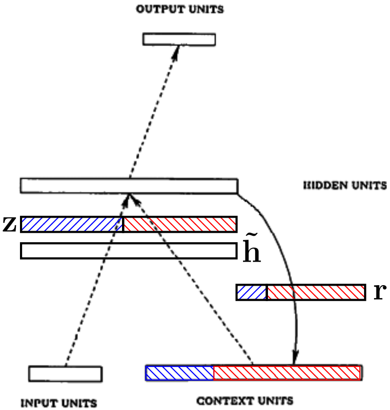
(Cho et al. 2014; Chung et al. 2015)

Gated recurrent neural networks

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r} \odot \mathbf{h}_{t-1}) + \mathbf{b})$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1} + \mathbf{b}_r)$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1} + \mathbf{b}_z)$$



(Cho et al. 2014; Chung et al. 2015)

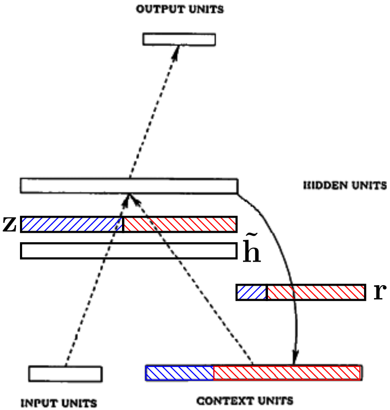
Gated recurrent neural networks

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r} \odot \mathbf{h}_{t-1}) + \mathbf{b})$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1} + \mathbf{b}_r)$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1} + \mathbf{b}_z)$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t$$



(Cho et al. 2014; Chung et al. 2015)