LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

CENTER FOR INFORMATION AND LANGUAGE PROCESSING

The past, present, and future of NLP from a linguistic perspective

Leonie Weissweiler
12.07.2022

Four questions

- What is the structure of language and how do we acquire it?
- What is the meaning of a word?
- Where in these debates are Transformers?
- Where do we go from here?

What is the structure of language and how do we acquire it?

English Past Tense

- Regular: need → needed
- Irregular: is → was, goes → went, comes → came etc.
- Three stages in child language acquisition

Verb Type	Early Verbs	Regular	Other Irregular	Novel
Stage 1	Correct	-	-	-
Stage 2	Regularized	Correct	Regularized	Regularized
Stage 3	Correct	Correct	Regularized	Regularized

→ Classic example for the debate: how do children learn this?

Rumelhart, D. E., & McClelland, J. L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing?.

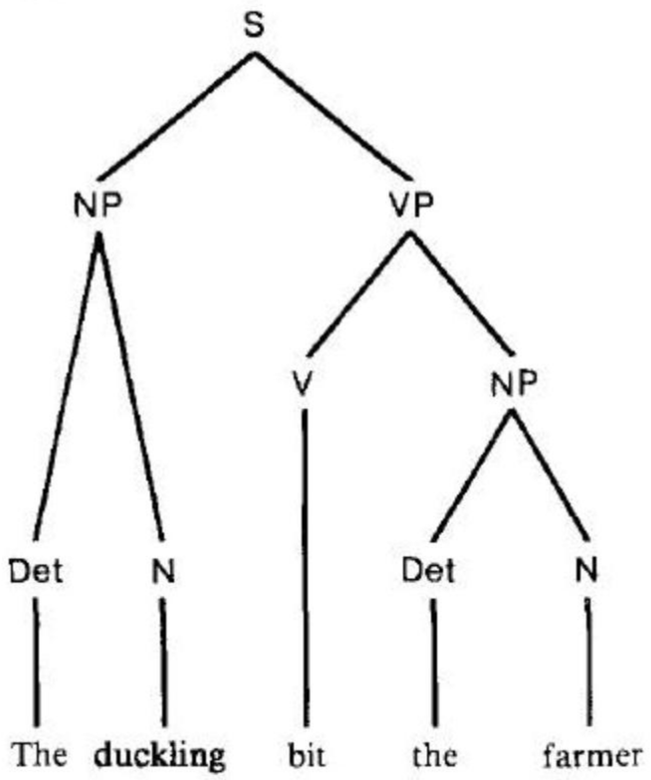
Chomsky 1957: Humans learn the rules of language

Language is “a system of rules that in some explicit and well-defined way assigns structural descriptions to sentences”

- S → NP + VP
- NP → Det + N

...

- Rule: Verb in Past Tense → Verb + ‘-ed’
- Lexicon: is → was, goes → went …



Chomsky, N. (1957). Syntactic structures.

Chomsky: Rules as an innate human bias

- Poverty of the Stimulus: Data that children are exposed to
  - is consistent with an infinite number of possible grammars
  - contains no negative feedback
  - is degenerate in terms of scope and quality
  - is different for each child

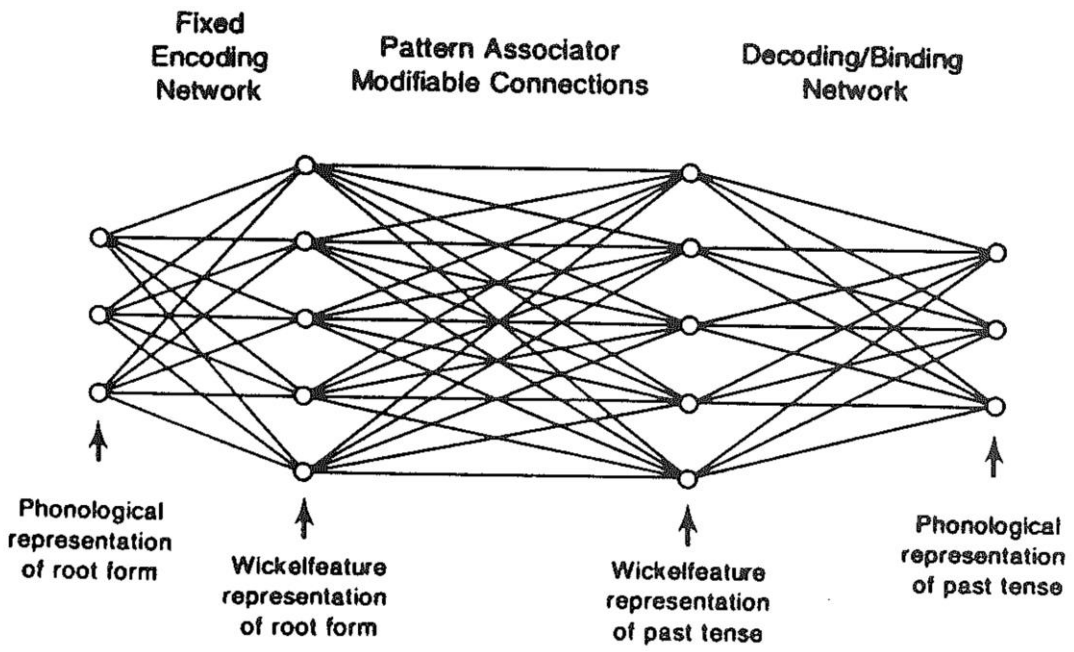
→ Language Acquisition Device

- Bias for tree-based grammar structure hardwired into the brain: Universal Grammar
- Contains options for language diversity that children simply choose from

Rumelhart, D. E., & McClelland, J. L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing?.

Rumelhart and McClelland 1986: Humans can learn with a Neural Network

- “Implicit knowledge of language may be stored in connections among simple processing units organized into networks”
- “Acquisition occurs by a simple process of adjusting connections between units”



→Past tense without explicit rules

→Joint handling of regular and irregular forms

→No separate lexicon for irregular verbs

Rumelhart, D. E., & McClelland, J. L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing?.

1988, Pinker & Prince point out issues with R&C's model

- R&M Model only correct in 67% of cases
- Uncharacteristic errors that mix forms, like eat → ated
- Over-irregularization, ping → pang

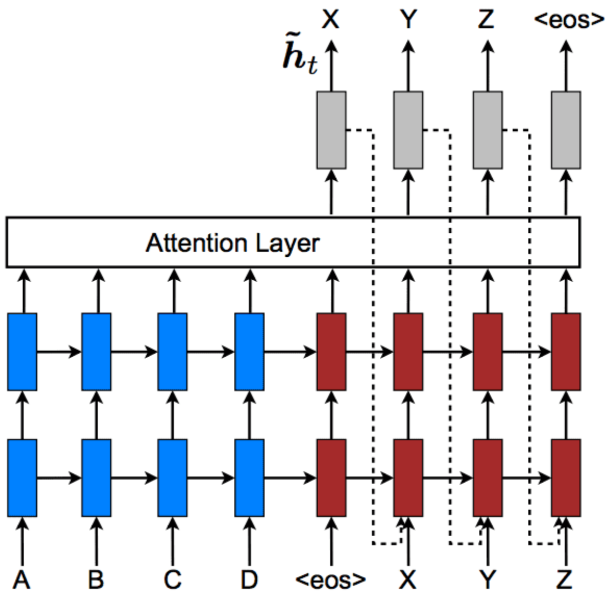
→ widespread skepticism towards NNs for modeling linguistic data and human cognition among linguists and cognitive scientists to this day

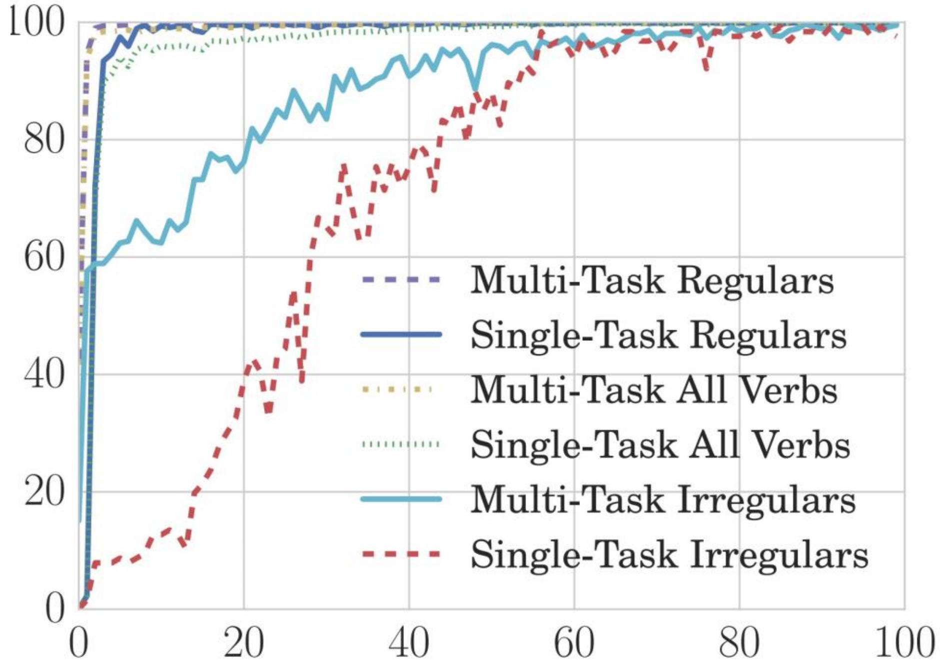
→ NLP likewise doesn't seriously use NNs for another few decades

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*.

2018, Kirov & Cotterell: Encoder-Decoder-Network

Two Recurrent Neural Networks with an attention mechanism



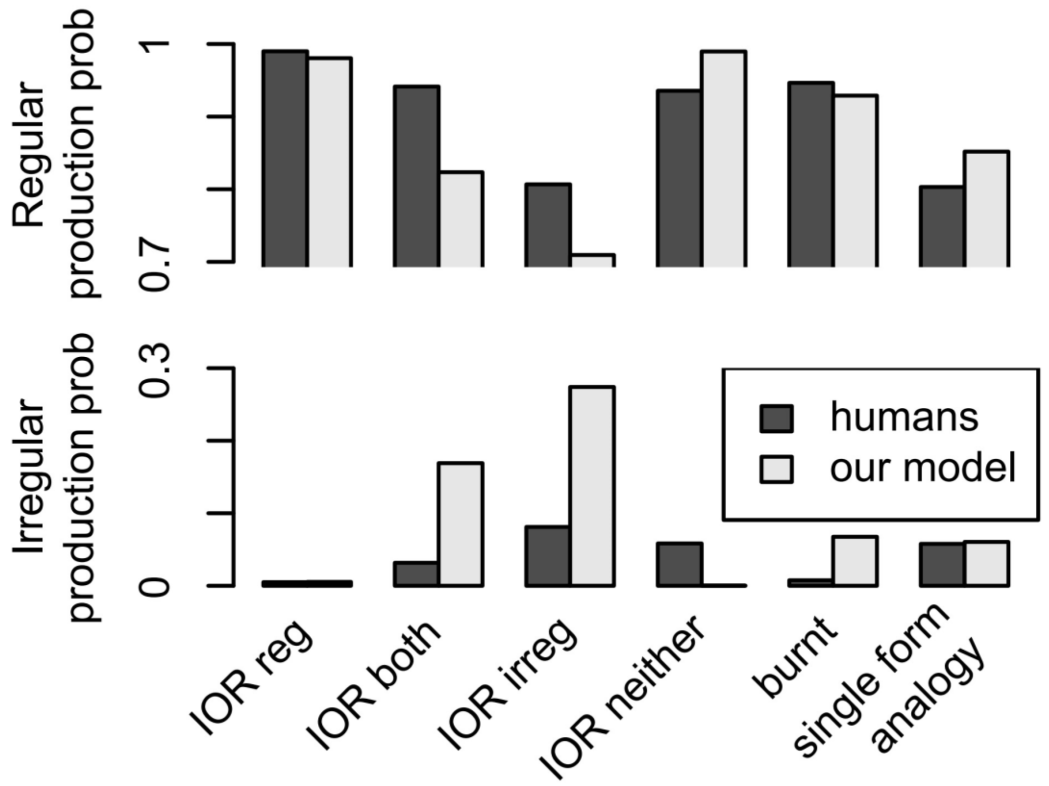


Form	Encoder-Decoder	MGL
Regular	**0.48**	0.35
Irregular	**0.45**	0.36

Kirov, C., & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. TACL

Corkery et al. 2019: Instability on Nonce Words

- Replicated K&C's accuracy on real verbs
- Instability over multiple runs of the model
- Overproduction of irregular forms for nonce verbs



→ The discussion remains open

Corkery, M., Matusevych, Y., & Goldwater, S. (2019). Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. ACL

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

What is the meaning of a word and how is it represented in the brain?

1950s: Distributional Semantics

"You shall know a word by the company it keeps" – Firth, 1957 Modes of Meaning

He filled the wampimuk with the substance, passed it around ad we all drunk some

vs.

We found a little, hairy wampimuk sleeping behind the tree.

→ What can we learn about wampimuks purely from context?

Firth, J. R. (1957). Modes of meaning, papers in linguistics.

1970s: Truth-Conditional Semantics

- The meaning of a sentence is the number of possible worlds in which this sentence is true
- → Evaluate truth condition of a sentence
- ‘If Socrates is a man and all men are mortal, then Socrates is mortal.’
- $[\text{Man}(a) \land \forall\, (\text{Man}(x) \rightarrow \text{mortal}(x)] \rightarrow \text{mortal}(a)$
- But:
  - Questions and commands
  - Modals (may, can, …)
  - Attitude (I believe that …)

1970s: Componential Analysis

Analyse the internal semantic structure of a word as composed of a number of distinct and minimal components of meaning

	Cat	Puma	Dog	Wolf
animate	+	+	+	+
domesticated	+	-	+	-
feline	+	+	-	-

Nida, Eugene A. (1979). Componential analysis of meaning : an introduction to semantic structures (2nd ed.)

Rosch 1973: Prototypes

- Categories do not have clear boundaries
- Humans agree on ‘how much’ something is a bird

→ Birdiness ranking

→ Fuzzy representation in the brain

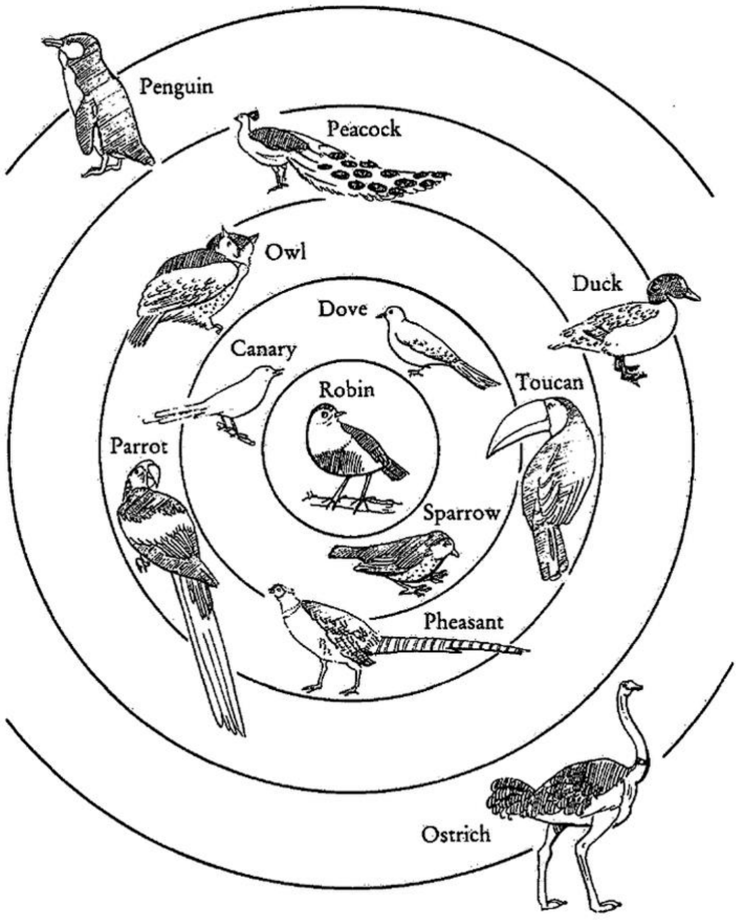


Figure . 1 Birdiness rankings

Rosch, E. H. (1973). Natural categories. *Cognitive psychology*.

1990s: Count-based Word Embeddings

Simply count how often words co-occur

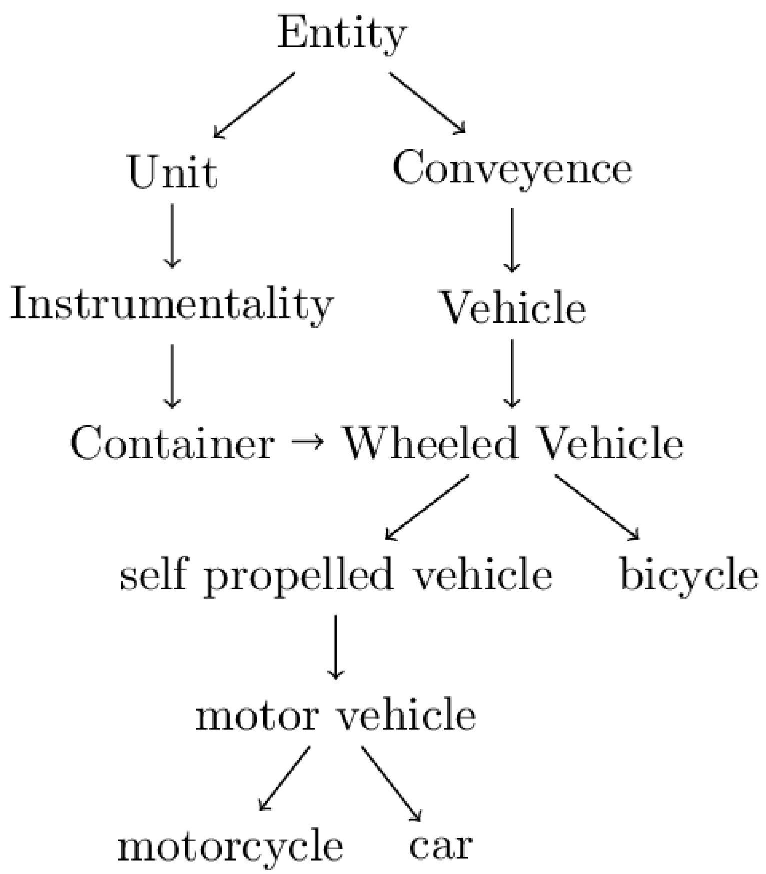
→ Incredibly sparse

The dog barked in the park.
The owner of the dog put him on the leash since he barked.

	leash	walk	run	owner	pet	bark
dog	3	5	2	5	3	2
cat	0	3	3	2	3	0
lion	0	3	2	0	1	0
light	0	0	0	0	0	0
bark	1	0	0	2	1	0
car	0	0	1	3	0	0

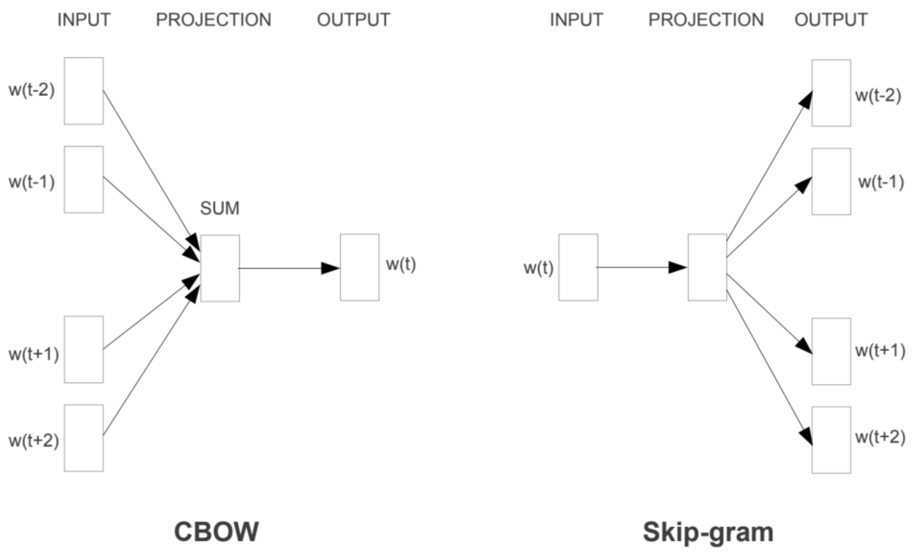
1995: WordNet

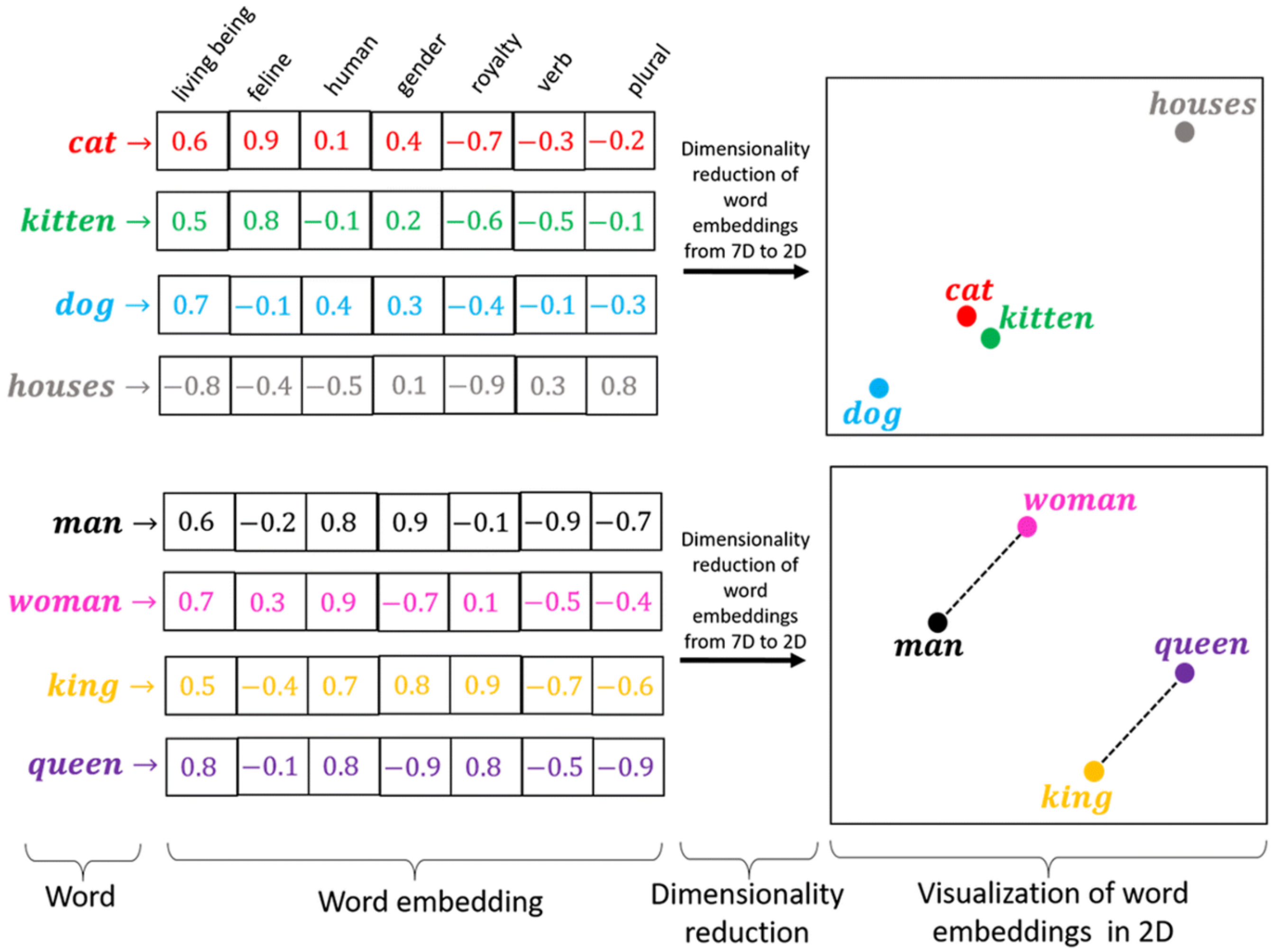
- Manually compiled
- Relations like synonymy, hyponymy, meronymy…
- But: struggles with abstract concepts



Miller, G. A. (1995). WordNet: a lexical database for English. ACM

2013: Trainable Word Embeddings





Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

https://medium.com/@hari4om/word-embedding-d816f643140

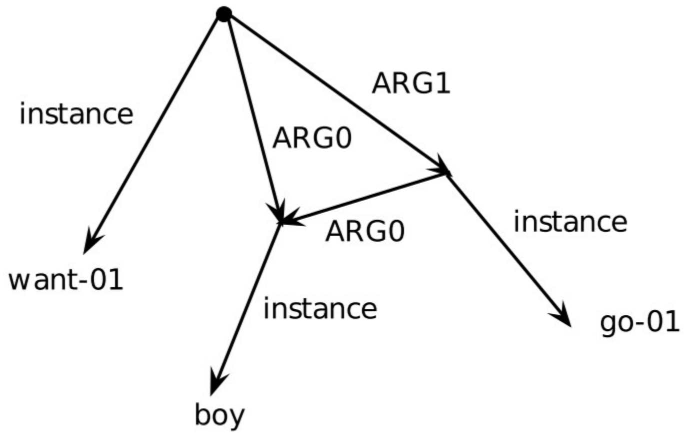
2013: Abstract Meaning Representation

The boy wants to go.

AMR format (based on PENMAN):

```
(w / want-01
  :arg0 (b / boy)
  :arg1 (g / go-01
          :arg0 b))
```

GRAPH format:



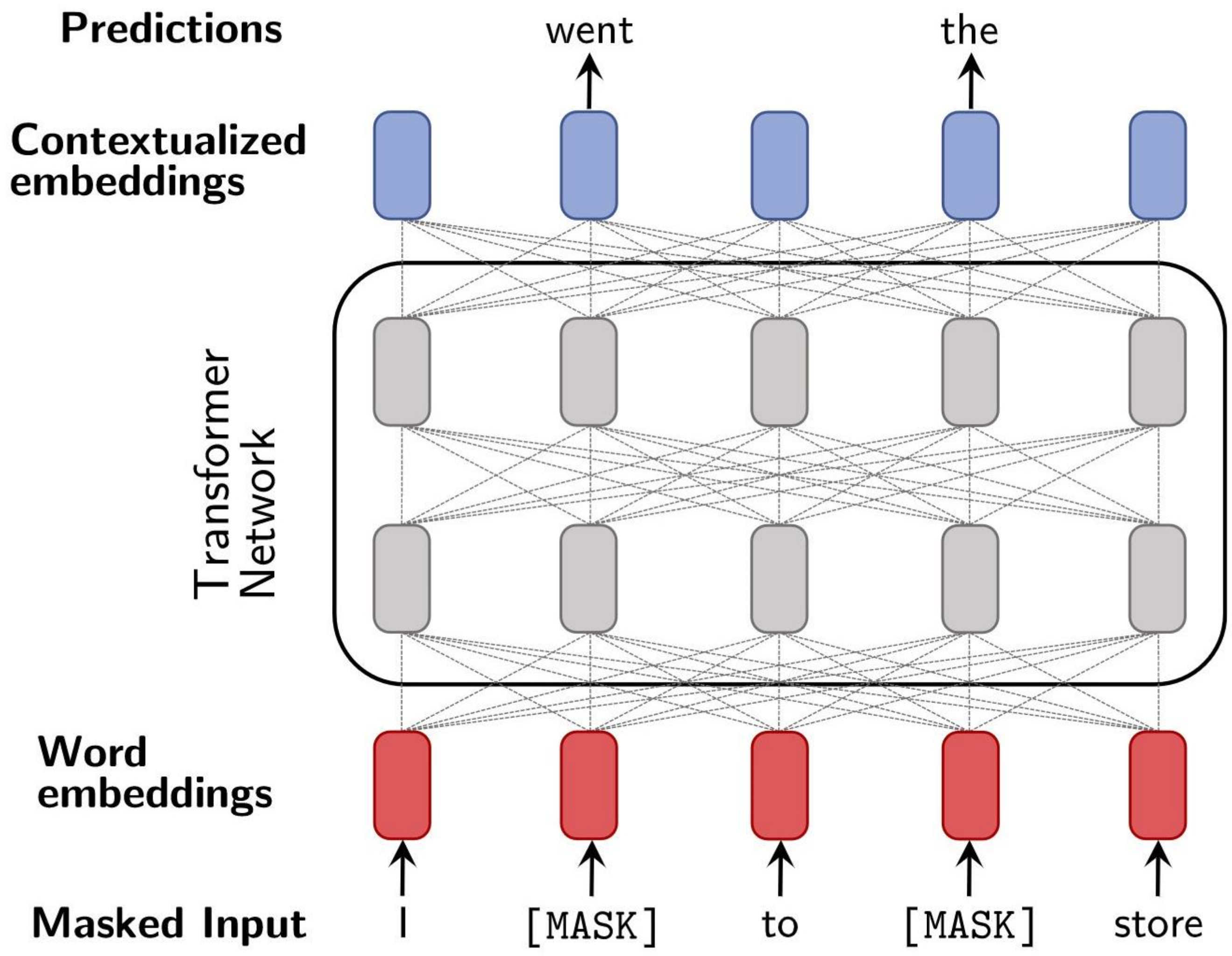
LOGIC format:

$\exists$ w, b, g:
instance(w, want-01) $\wedge$ instance(g, go-01) $\wedge$
instance(b, boy) $\wedge$ arg0(w, b) $\wedge$
arg1(w, g) $\wedge$ arg0(g, b)

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., ... & Schneider, N. (2013, August). ACL / LAW

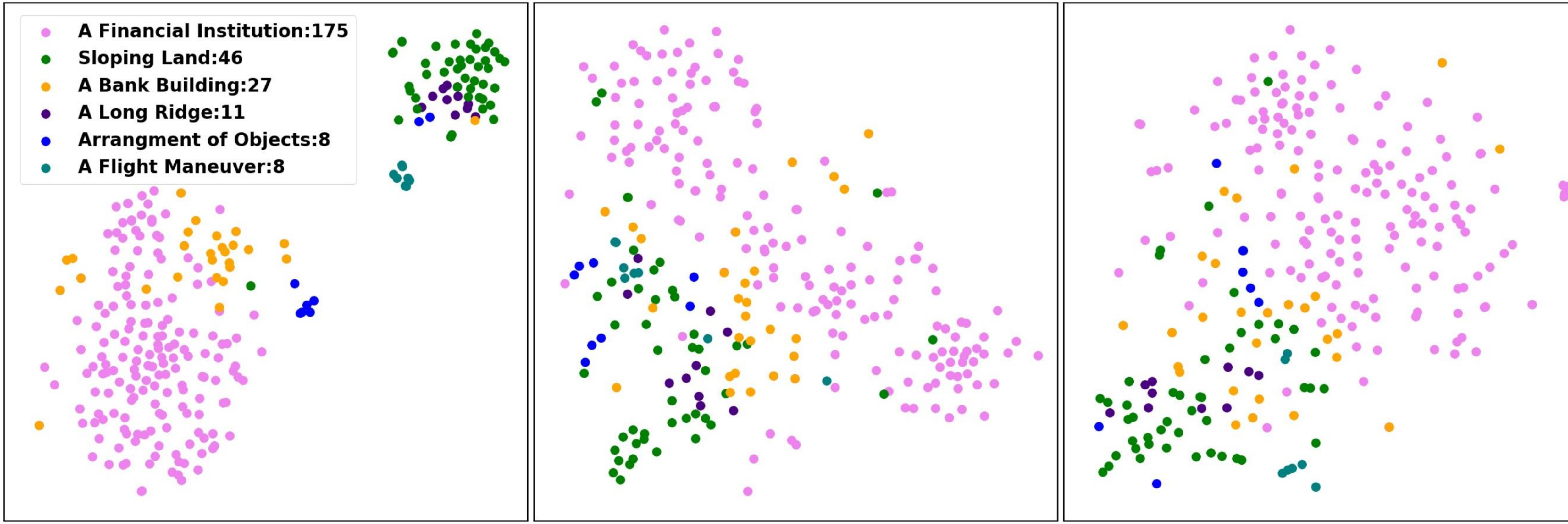
Where in these debates are Transformers?

Transformers: the Victory of Connectionism?



Hewitt, J., & Manning, C. D. (2019, June). A structural probe for finding syntax in word representations. NAACL

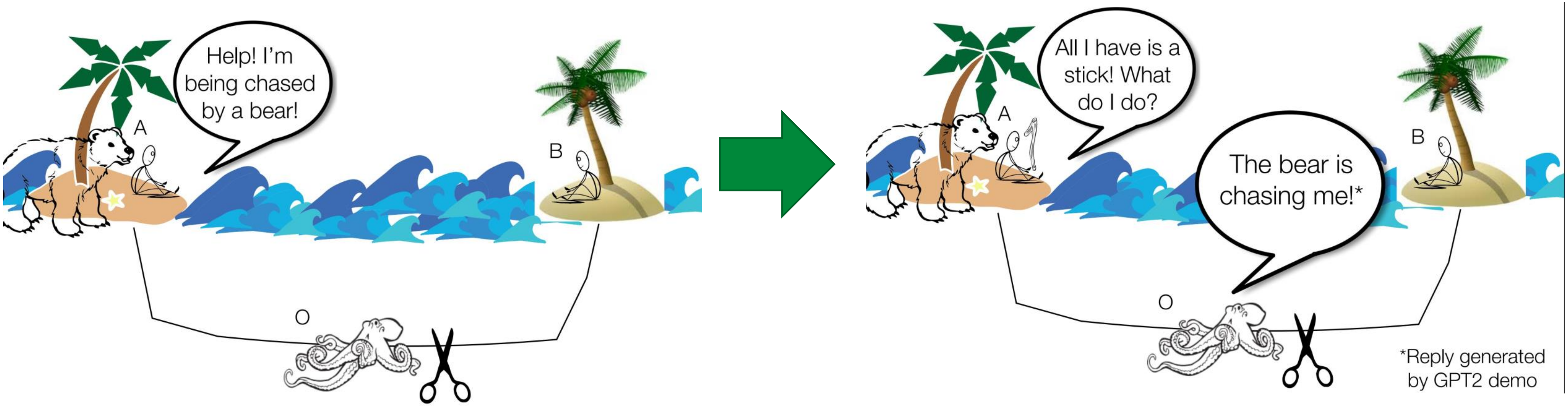
Contextual Embeddings



(a) BERT (b) Flair (c) ELMo

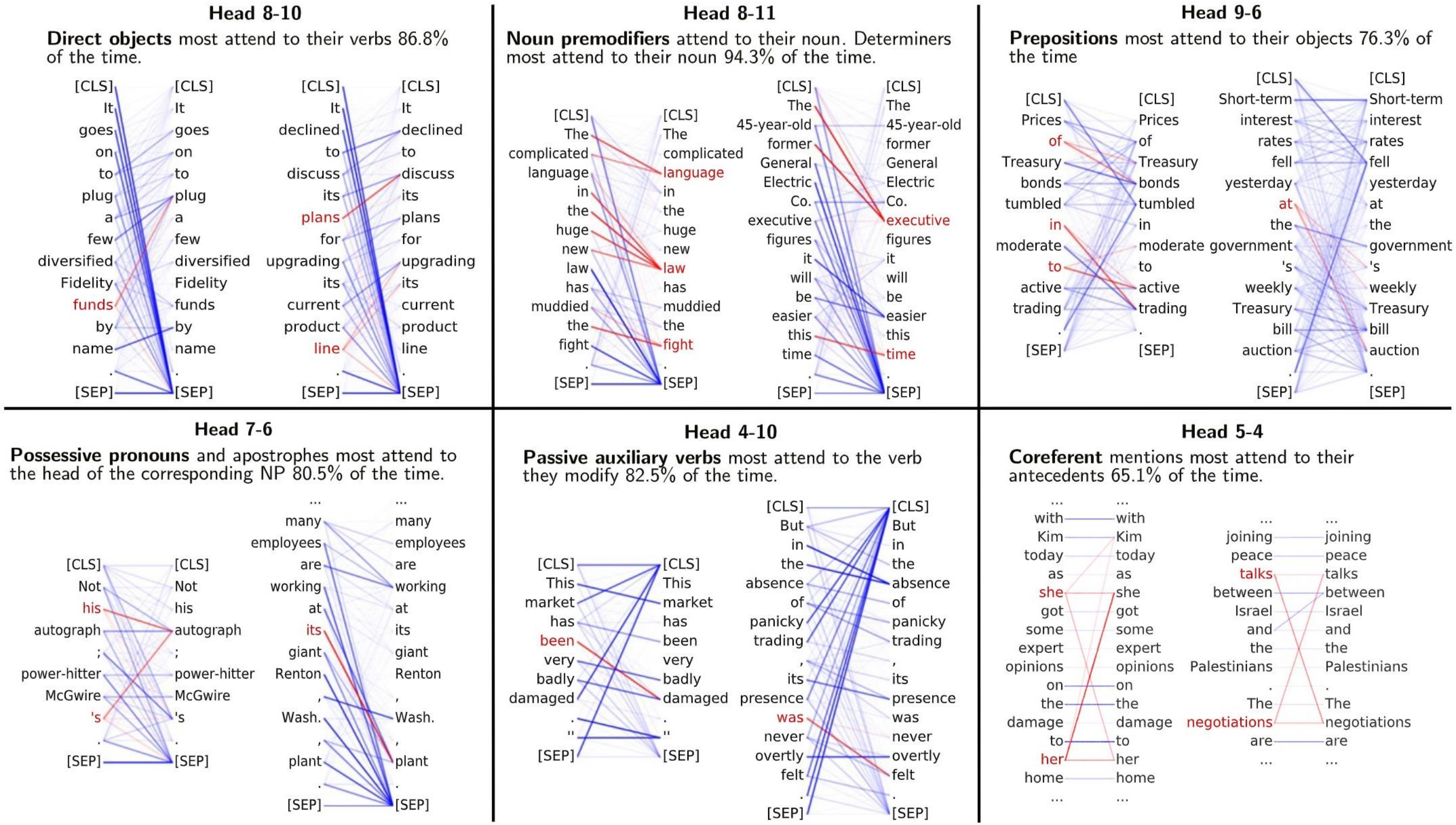
Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings.

Meaning purely from text?



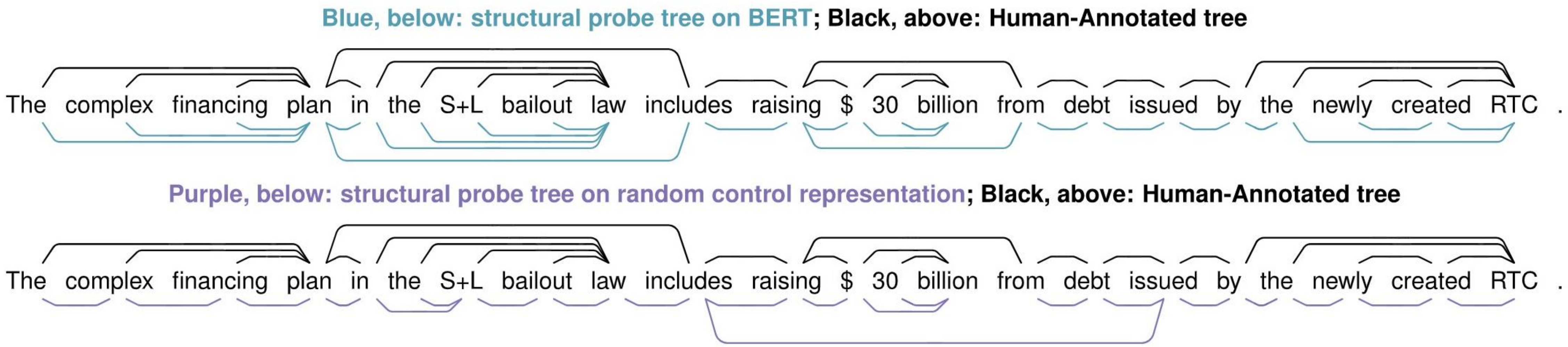
Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. ACL

Probing for Dependency Syntax



Hewitt, J., & Manning, C. D. (2019, June). A structural probe for finding syntax in word representations. NAACL

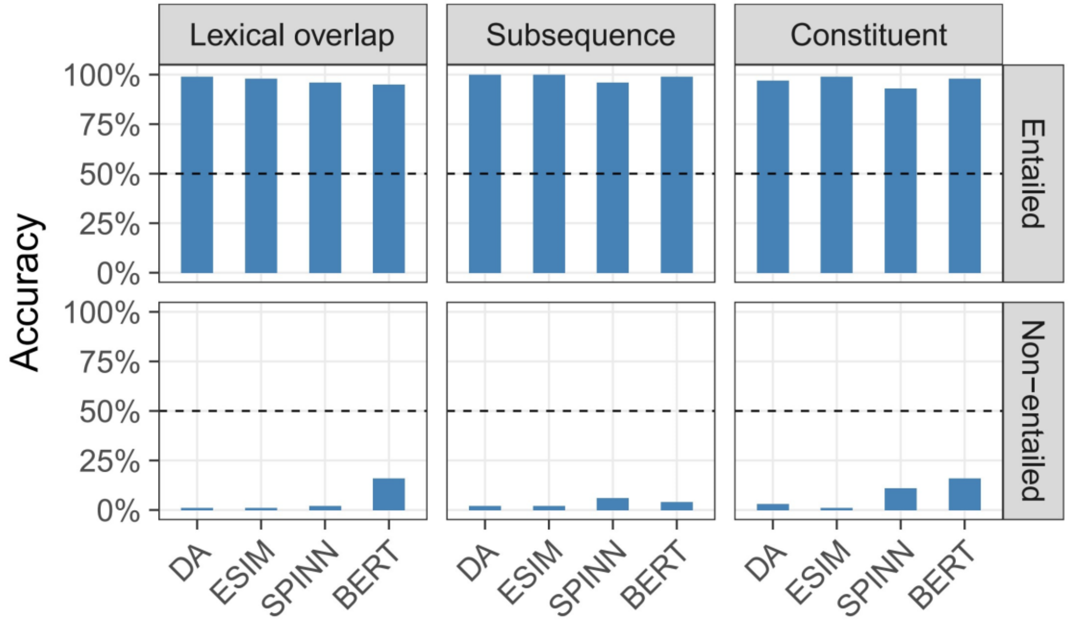
Structural probes



Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *PNAS*

Right for the wrong reasons?

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.



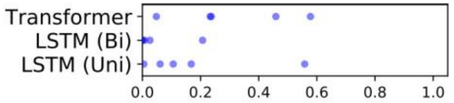
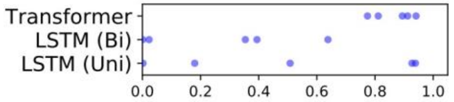
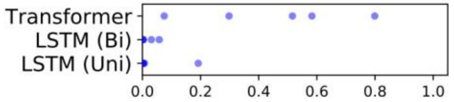
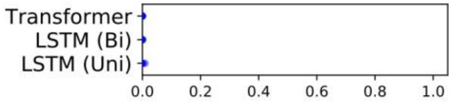
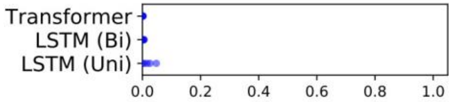
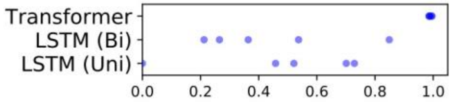
McCoy, T., Pavlick, E., & Linzen, T. (2019, July). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ACL

Compositionality

- COGS

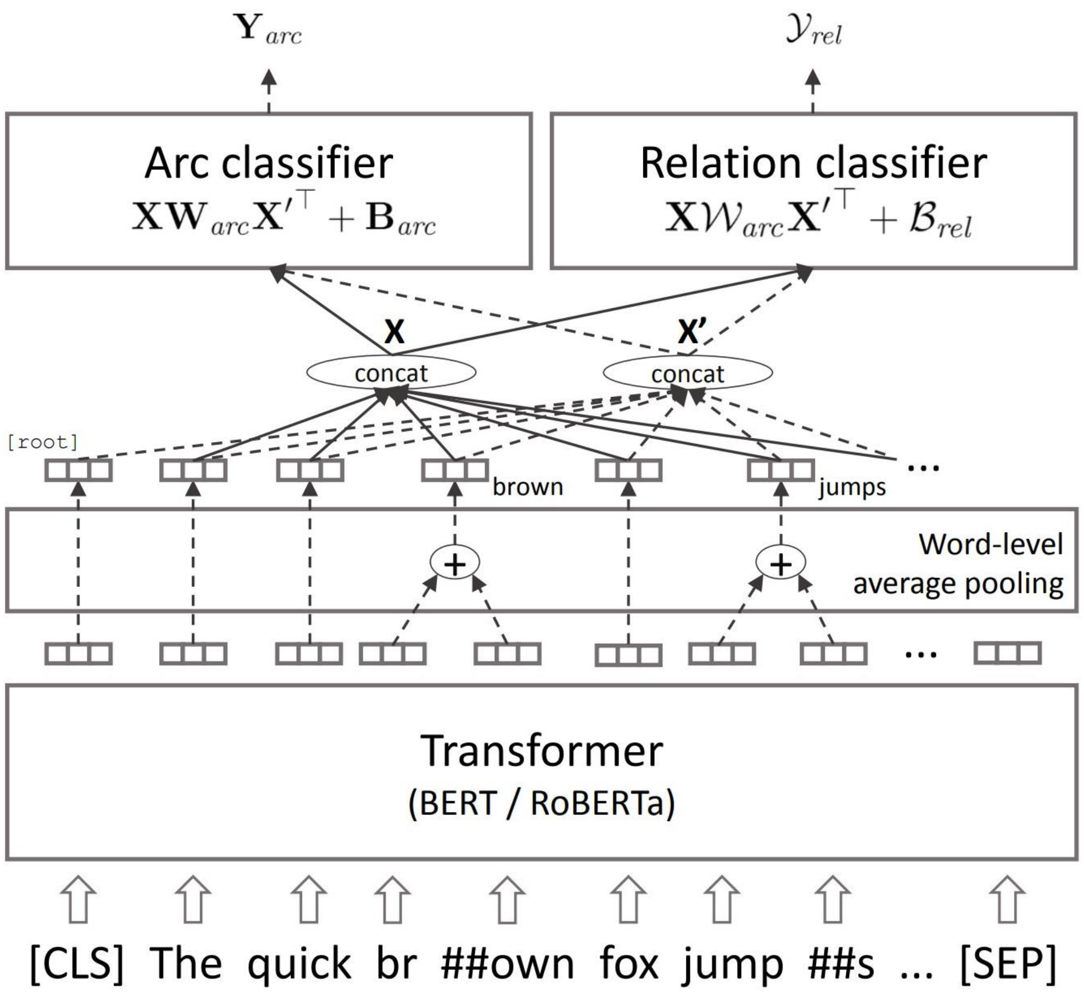
antic

Interpr

Case	Training	Generalization	Accuracy Distribution
Subject → Object (common noun)	*Subject* A **hedgehog** ate the cake.	*Object* The baby liked the **hedgehog**.	
Object → Subject (common noun)	*Object* Henry liked a **cockroach**.	*Subject* The **cockroach** ate the bat.	
Object → Subject (proper noun)	*Object* Mary saw **Charlie**.	*Subject* **Charlie** ate a donut.	
Primitive → Object (proper noun)	*Primitive* **Paula**	*Object* The child helped **Paula**.	
Depth generalization: PP modifiers	*Depth 2* Ava saw the ball **in the bottle on the table**.	*Depth 3* Ava saw the ball **in the bottle on the table on the floor**.	
Active → Passive	*Active* Emma **blessed** William.	*Passive* A child **was blessed**.	

Kim, N., & Linzen, T. (2020, November). COGS: A compositional generalization challenge based on semantic interpretation. EMNLP

Does injecting structure help?



Glavaš, G., & Vulić, I. (2021, April). Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation. EMNLP

Some Questions for Discussion

- For Linguistics: does the success of Neural Networks count in favour of connectionist modeling? What do the improvements with ever larger data mean for the Poverty of the Stimulus?
- For NLP: how do we want our models to develop? Are we going to bring formal syntax or formal semantics back into Transformer models?
- For ML: what biases are large neural networks developing?

Discussion

How should LMs learn language?

Leonie Weissweiler
Oettingenstraße 67· 80538 München
weissweiler@cis.lmu.de · www.cis.lmu.de/~weissweiler · @laweissweiler