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

<table>
<thead>
<tr>
<th>Verb Type</th>
<th>Early Verbs</th>
<th>Regular</th>
<th>Other Irregular</th>
<th>Novel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>Correct</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Stage 2</td>
<td>Regularized</td>
<td>Correct</td>
<td>Regularized</td>
<td>Regularized</td>
</tr>
<tr>
<td>Stage 3</td>
<td>Correct</td>
<td>Correct</td>
<td>Regularized</td>
<td>Regularized</td>
</tr>
</tbody>
</table>

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

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”

- $S \rightarrow NP + VP$
- $NP \rightarrow \text{Det} + \text{N}$

... 

- Rule: Verb in Past Tense $\rightarrow$ Verb + ‘-ed’
- Lexicon: is $\rightarrow$ was, goes $\rightarrow$ went ...
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 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
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

a parallel distributed processing model of language acquisition. Cognition.
Two Recurrent Neural Networks with an attention mechanism

<table>
<thead>
<tr>
<th>Form</th>
<th>Encoder-Decoder</th>
<th>MGL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular</td>
<td>0.48</td>
<td>0.35</td>
</tr>
<tr>
<td>Irregular</td>
<td>0.45</td>
<td>0.36</td>
</tr>
</tbody>
</table>
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
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.
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

<table>
<thead>
<tr>
<th></th>
<th>Cat</th>
<th>Puma</th>
<th>Dog</th>
<th>Wolf</th>
</tr>
</thead>
<tbody>
<tr>
<td>animate</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>domesticated</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>feline</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

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.

<table>
<thead>
<tr>
<th></th>
<th>leash</th>
<th>walk</th>
<th>run</th>
<th>owner</th>
<th>pet</th>
<th>bark</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>cat</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>lion</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>light</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bark</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>car</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
1995: WordNet

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

2013: Trainable Word Embeddings


https://medium.com/@hari4om/word-embedding-d816f643140
2013: Abstract Meaning Representation

The boy wants to go.

**AMR format** (based on PENMAN):

\[(w / \text{want-01})\]
\[:\text{arg}\text{0} (b / \text{boy})\]
\[:\text{arg}\text{1} (g / \text{go-01})\]
\[:\text{arg}\text{0} b)\]

**GRAPH format:**

**LOGIC format:**

\[
\exists w, b, g:
\text{instance}(w, \text{want-01}) \land \text{instance}(g, \text{go-01}) \land \text{instance}(b, \text{boy}) \land \text{arg0}(w, b) \land \text{arg1}(w, g) \land \text{arg0}(g, b)
\]
Where in these debates are Transformers?
Transformers: the Victory of Connectionism?

Contextual 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

Structural probes

Blue, below: structural probe tree on BERT; Black, above: Human-Annotated tree

The complex financing plan in the S+L bailout law includes raising $30 billion from debt issued by the newly created RTC.

Purple, below: structural probe tree on random control representation; Black, above: Human-Annotated tree

The complex financing plan in the S+L bailout law includes raising $30 billion from debt issued by the newly created RTC.
## Right for the wrong reasons?

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical overlap</td>
<td>Assume that a premise entails all hypotheses constructed from words in the premise</td>
<td>The doctor was <strong>paid</strong> by the actor.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>→ The doctor paid the actor.</td>
</tr>
<tr>
<td>Subsequence</td>
<td>Assume that a premise entails all of its contiguous subsequences.</td>
<td>The doctor near the actor <strong>danced</strong>.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>→ The actor danced.</td>
</tr>
<tr>
<td>Constituent</td>
<td>Assume that a premise entails all complete subtrees in its parse tree.</td>
<td>If the artist <strong>slept</strong>, the actor ran.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>→ The artist slept.</td>
</tr>
</tbody>
</table>

![Accuracy Graphs](image1)

### COGS: A Compositional Generalization Challenge Based on Semantic Interpretation

#### Case

<table>
<thead>
<tr>
<th>Case</th>
<th>Training</th>
<th>Generalization</th>
<th>Accuracy Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject → Object (common noun)</td>
<td>Subject A <em>hedgehog</em> ate the cake.</td>
<td>Object The baby liked the <em>hedgehog</em>.</td>
<td><img src="image" alt="Accuracy Distribution" /></td>
</tr>
<tr>
<td>Object → Subject (common noun)</td>
<td>Object Henry liked a <em>cockroach</em>.</td>
<td>Subject The <em>cockroach</em> ate the bat.</td>
<td><img src="image" alt="Accuracy Distribution" /></td>
</tr>
<tr>
<td>Object → Subject (proper noun)</td>
<td>Object Mary saw <em>Charlie</em>.</td>
<td>Subject <em>Charlie</em> ate a donut.</td>
<td><img src="image" alt="Accuracy Distribution" /></td>
</tr>
<tr>
<td>Primitive → Object (proper noun)</td>
<td>Primitive <em>Paula</em></td>
<td>Object The child helped <em>Paula</em>.</td>
<td><img src="image" alt="Accuracy Distribution" /></td>
</tr>
<tr>
<td>Depth generalization: PP modifiers</td>
<td>Depth 2 Ava saw the ball in the bottle on the table.</td>
<td>Depth 3 Ava saw the ball in the bottle on the table on the floor.</td>
<td><img src="image" alt="Accuracy Distribution" /></td>
</tr>
<tr>
<td>Active → Passive</td>
<td>Active Emma <em>blessed</em> William.</td>
<td>Passive A child was <em>blessed</em>.</td>
<td><img src="image" alt="Accuracy Distribution" /></td>
</tr>
</tbody>
</table>

Does injecting structure help?

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?