LING-UA 1 *Language*

Language Models & Linguistics

Jackson Petty

Department of Linguistics

SILV #104

2:00pm, 9 October 2025

What does it mean to know a language?

- Humans can…
  - produce unbounded novel constructions
  - discern grammaticality on the basis of invisible structure
  - assign & interpret meaning to utterances (∼compositionally)
  - acquire these skills from “naturalistic” training data
- This argues for Universal Grammar
  - Human language faculty “bakes in” many assumptions about how Language works
  - enables “rapid” acquisition from small amounts of input

“Knowing a Language” is really valuable!

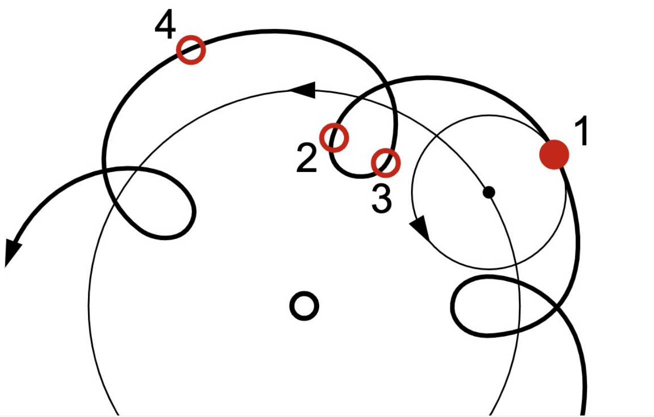
- Lots of tasks which require linguistic knowledge:
  - Translation
  - Document summarization
  - Question-answering
  - Automated writing
- Could we use our understanding of UG to build a computational system that uses natural language?

  Empirically: not very well :(

The Old Approach to NLP

- How Natural Language Processing used to work:
  - Build parser-like systems that function like akin to how human language processing (must) work
  - Every time you encounter an edge case?
    Add explicit rules to deal with it
- This approach has some problems:
  - It's *very* brittle, and human language has *lots* of…fuzziness
    - misspelling, dialectal variation, slang, emojis, non-linguistic data (numbers, pictures)
  - The complexity of the system grows monotonically
  - Requires highly-annotated data to identify invisible structure

NLP, c. 2010: "Just add more epicycles!"



A new approach

- What if we got rid of $\sim$*every* structural assumption?
- The **only** thing we care about: *predicting the next word*

The Language Model Game

[BOS] ___

[BOS] an apple boy eats the .

The Language Model Game

[BOS] ___

The Language Model Game

[BOS] The ___

[BOS]	an	apple	boy	eats	the	.

The Language Model Game

[BOS] The ___

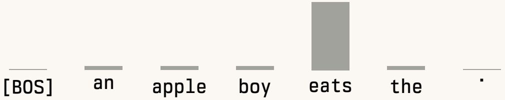
The Language Model Game

[BOS] The boy ___

[BOS]	an	apple	boy	eats	the	.

The Language Model Game

[BOS] The boy ___



The Language Model Game

[BOS] The boy eats ___

[BOS]	an	apple	boy	eats	the	.

The Language Model Game

[BOS] The boy eats ___

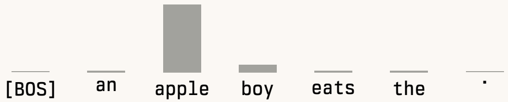
The Language Model Game

[BOS] The boy eats an ___

[BOS]   an   apple   boy   eats   the   .

The Language Model Game

[BOS] The boy eats an ___



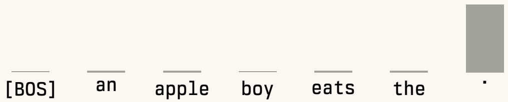
The Language Model Game

[BOS] The boy eats an apple ___

[BOS]	an	apple	boy	eats	the	.

The Language Model Game

[BOS] The boy eats an apple ___



Language Models

- **Language models** work by predicting "next-word" probabilities
  - a function which tells you the probabilities of new words given a partial sentence
  - they make *no assumptions* about "structure" or "meaning" or anything like that
- How do we estimate these probabilities?
  1. Get a bunch of data
  2. *Compress data into a probabilistic model* [magic happens]
  3. Use this compressed model to predict probabilities for new text

Two Kinds of Language Model

- ***n*-gram Model** ($n = 2, 3, \ldots$)
  - Take every possible 2-word (3-word) pair, eg, ("the", "boy")
  - Measure how of often that pair appears in a corpus
  - This becomes your estimate of the probability for `the boy`!
- **Neural language model**
  - Stack a bunch of matrices with random values in them
    = *neural network* (greatly simplified)
  - Use this stack to predict sequences from a corpus (very bad!)
  - Measure how wrong those predictions are, apply calculus to change the values in the matrix
  - Repeat trillions of times

Why Language Models?

- They work *shockingly well*
- They don’t need highly-annotated data (= *self-supervision*)
- Quality scales with amount of data and compute (highly fungible)
- Handle the “fuzziness” of language (more) gracefully

What does this mean?

- Do LMs “know” the same thing as people?
- Are LMs good models of human language acquisition?
- Why did this probabilistic approach succeed where the linguistically-informed, rule-based approach failed?

What does this mean?

- Do LMs “know” the same thing as people?
- Are LMs good models of human language acquisition?
- Why did this probabilistic approach succeed where the linguistically-informed, rule-based approach failed?

Do LMs "know" the same thing as people?

- LMs seem to generate plausible text
- How can we compare LM-text to human-generated text?
  - UG makes claims about how humans *generalize* to new data
  - $\longrightarrow$ does an LM make the same generalizations?
  - Measure this by looking at specific linguistic phenomena

Example: Long-Distance Agreement (Linzen, Dupoux & Goldberg 2021)

- **Central claim of UG:** language operates on (hierarchical) structure, not (linear) strings
  - → subject-verb agreement follows hierarchy, not a linear rule
  - The **length** of the *forewings* (is/*are)
- **Setup:**
  - train an LM on (simple, ambiguous) sentences
  - test on (complex, disambiguating)
  - *scale # of distractors:*

    The **key** to the *cabinets* by the *doors* ___
- **Results:**
  - Neural LMs do quite well on "common" structures, less well on "uncommon" ones

What does this mean?

- Do LMs “know” the same thing as people?
- Are LMs good models of human language acquisition?
- Why did this probabilistic approach succeed where the linguistically-informed, rule-based approach failed?

Are LMs good models of human language acquisition? (BabyLM)

- Humans learn from *shockingly* little data
  → Chomsky's *Poverty of the Stimulus* argument
- LMs are *good*, but need *tons* of data:

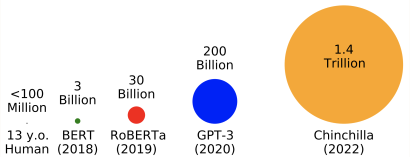


Figure: ~7 orders of magnitude between humans and 'good' LLMs