Experimental and Computational Methods in Linguistic Research

Spring 2025

Instructor: Sanghee Kim

Week 6

Agenda

- Preprocessing PCIbex data
- Plotting line graph
- Comparison between human reading times and model output

- Preprocessing PCIbex data
- Plotting line graph

Number agreement attraction effect

(a) The key to the cabinet was rusty.
(b) The key to the cabinets was rusty.
(c) The key to the cabinet were rusty.
(d) The key to the cabinets were rusty.

Number agreement attraction effect

- Prediction on the reading time @was/were (+1)?

(a) **The key** to the cabinet **was** rusty.

(b) **The key** to the cabinets **was** rusty.

(c) ***The key** to the cabinet **were** rusty.

(d) ***The key** to the cabinets **were** rusty.

- (The most common pattern:) (c) > (d) > (a) ≈ (b)

Understanding reading times

- Why do we see such reading time differences?

The debate

- Memory?
- Expectation?

“Rick is starting a tornado garden”

“Rick is starting a tornado garden”

“Rick is starting at a NATO garden”

“Rick is starting a tomato garden”

“Rickets art innate omit a carton”

“Rick is starting a tornado garden”

“Rick is starting at a NATO garden”

“Rick is starting a tomato garden”

“Rickets art innate omit a carton”

Discussion

- How did you know? (Where did your assessment come from?)
- Do humans assign probabilities to strings of words?

Probabilities for language models

- Sandy went to the bakery and bought ???.

- To make bread, you at least need water, salt, and ???.

Probabilities for language models

- Sandy went to the bakery and bought ???.
  - How likely is it to see *bread?*
  - How likely is it to see *pajamas?*

- To make bread, you at least need water, salt, and ???.
  - How likely is it to see *flour*?
  - How likely is it to see *glue*?

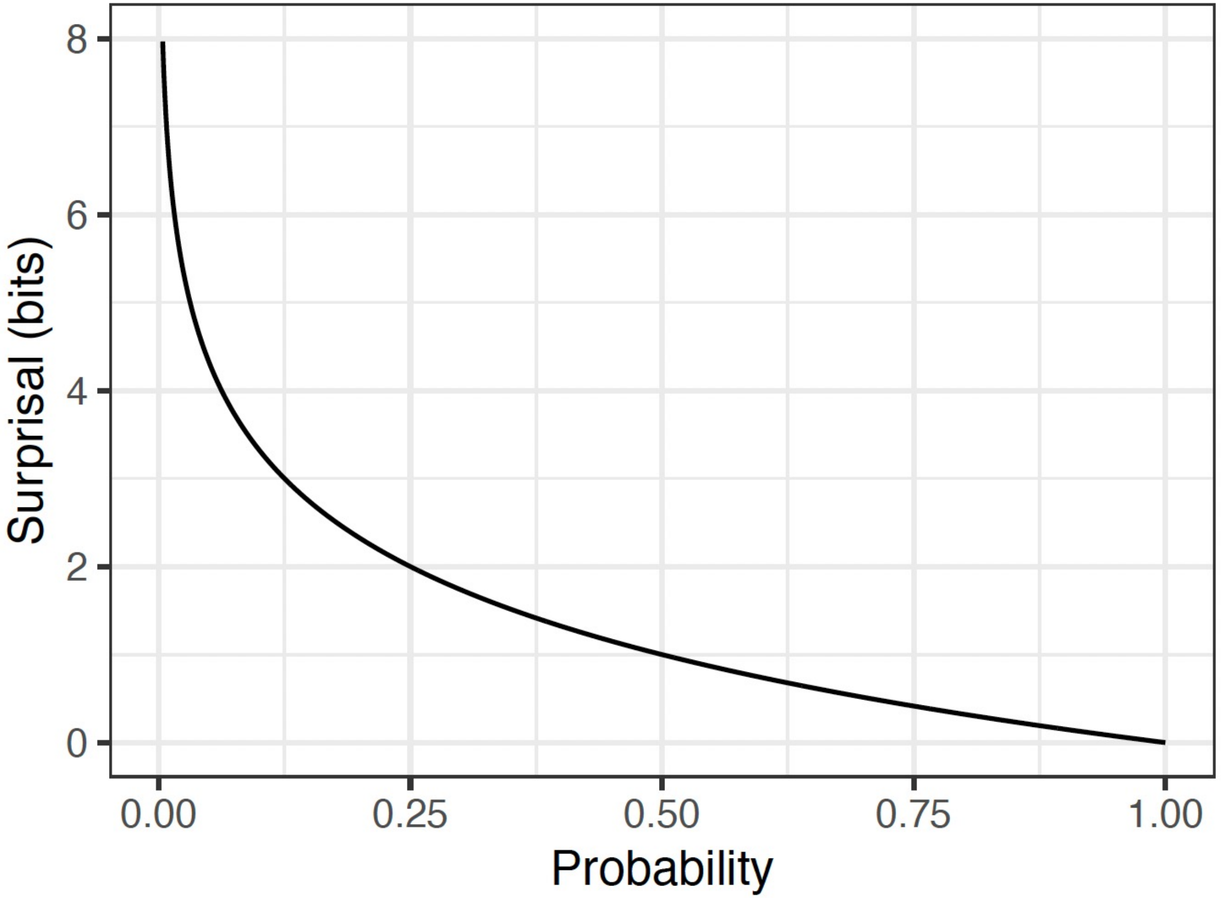
Informativity

- Sandy went to the bakery and bought <u>???</u>.

- How informative is ‘bread’ compared to ‘pajamas’?
- How surprised are you to see ‘bread’ compared to ‘pajamas’?

- Hypothesis: a word's difficulty is its *surprisal* in context:

$$\text{Surprisal}(w_i) \equiv \log \frac{1}{P(w_i|\text{CONTEXT})}$$



(Shannon, 1948: a basic quantity from information theory!)

Surprisal & Psycholinguistics

- In addition to measuring the average information for a language, we can of course measure the **information conveyed by any given linguistic unit** (e.g. phoneme, word, utterance) in context. This is often called *surprisal*:

$$Surprisal(x) = \log_2 \frac{1}{P(x \mid context)}$$

- **Surprisal will be high**, when *x* has a low conditional probability, and **low**, when *x* has a high probability.

- Claim: **Cognitive effort** required to process a word is **proportional** to its **surprisal** (Hale, 2001).

Computing Surprisal

$$\text{Surprisal}_{k+1} = -\log P(w_{k+1} \mid w_1 \ldots w_k)$$

- There are various ways we can compute surprisal from different kinds of underlying probabilistic language models
- N-gram surprisal:

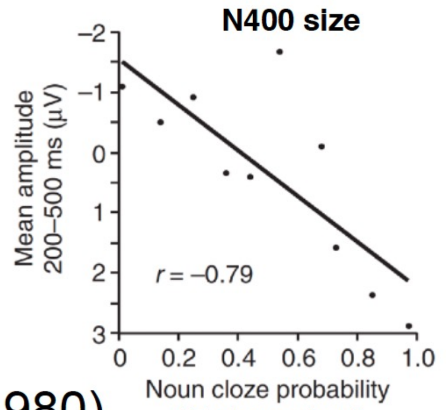
$$\text{Surprisal}(w_{k+1}) = -\log_2 p(w_{k+1} \mid w_{k-2}, w_{k-1}, w_k)$$

Surprisal as an index of real-time processing load

- Let a word's difficulty be its *surprisal* given its context:

$$\text{Surprisal}(w_i) \equiv \log \frac{1}{P(w_i|\text{CONTEXT})}$$
$$\left[\approx \log \frac{1}{P(w_i|w_{1\ldots i-1})}\right]$$

- Captures the *expectation* intuition: the more we expect an event, the easier it is to process
  - Brains are prediction engines!



(Federmeier & Kutas, 1999)

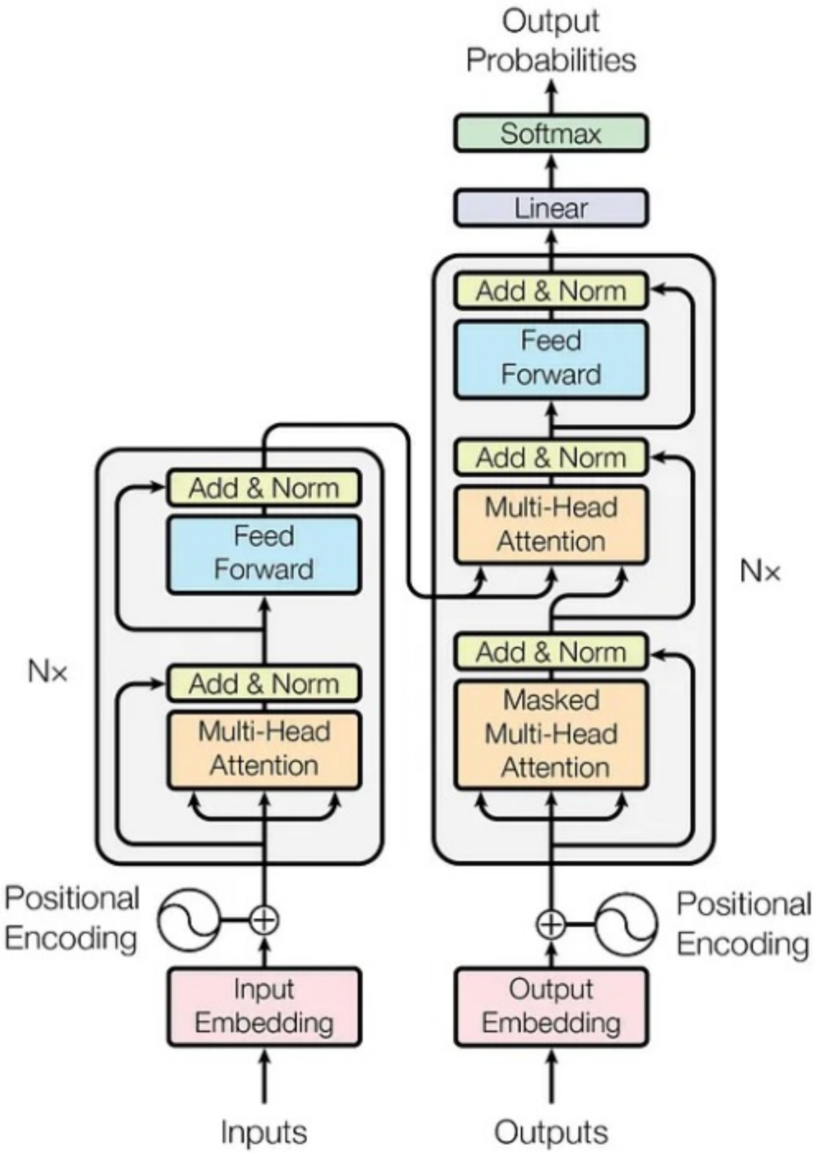
- Predictable words are:
  - read faster (Ehrlich & Rayner, 1981)
  - have distinctive EEG responses (Kutas & Hillyard 1980)

- with a language model that captures syntactic structure, we can get GRAMMATICAL EXPECTATIONS

(Hale, 2001, NAACL; Levy, 2008, *Cognition*)

BERT

Encoder



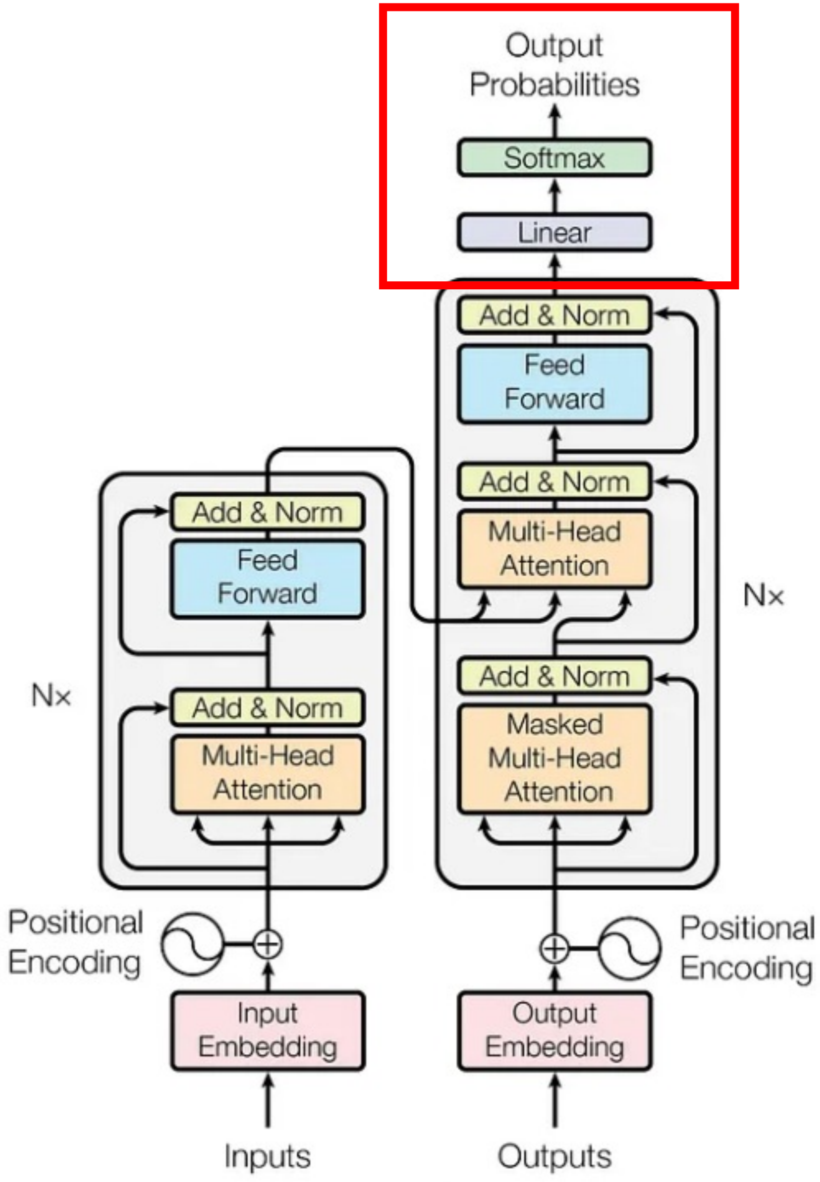
Transformer Architecture

GPT

Decoder

The approach (similar to Arehalli & Linzen, 2020):

- Obtain model surprisal at the critical word
- Compare it with human reading time results



Transformer Architecture