Surprisal from Large Language Models

AriaRay Brown Julius Steuer

17.05.2024

TaCoS

Goals of this Tutorial

1. Learn to calculate surprisal with our toolkit
2. Learn to use surprisal for psycholinguistic research
3. Learn to calculate surprisal with a language model

Processing Difficulty ~ Surprisal

- Surprisal theory: processing difficulty of a word in context is proportional to its negative log-probability (Hale 2001, Levy 2008)

$$\text{processing difficulty} \approx -\log_2 p(word|context)$$

- If we are interested in the processing difficulty of a word *n* given its context, we calculate its negative log probability given the *k* preceding words

$$\mathbf{surp}(w_n|w_{n-k}, ..., w_{n-1}) = -\log_2 p(w_n|w_{n-k}, ..., w_{n-1})$$

Surprisal from Language Models

- Instead of conditioning on the *n* preceding words, next word predictions are conditioned on the hidden state of transformer models

$$h_{n-1} = f_{\theta_{TF}}(w_{n-k}, ..., w_{n-1})$$

$$p(w_n|h_{n-1}) = softmax(f_{\theta_{LM}}(h_{n-1}))$$

- θ_{LM} can be an n-gram model, LSTM, transformer…

Predictions of Surprisal Theory

- Language model surprisal has been used successfully for reading time prediction (Smith & Levy 2013, Shain et al. 2022)

$$\text{Surp}(\underline{\text{would be changed}} \mid \text{The employees understood that the contract})$$

$$\propto$$

$$\text{RT}(\underline{\text{would be changed}} \mid \text{The employees understood that the contract})$$

- Surprisal theory should be able to **fully explain** reading times.

Reading Time fit with LME Models

- LME = Linear Mixed Effects (model), regression with more than one predictor
- Regression

Reading Time ~ Surprisal

- LME

Reading Time ~ Surprisal(w) + Frequency(w) + Length(w) + 1|w + ...

LME Goodness of Fit

- LMEs are compared via their log-likelihoods (LL)
  - How likely is my data given the fitted LME?
- Compare LME with surprisal as a predictor to a base model
  - LL_surp = Reading Time ~ Surprisal(w) + Frequency(w) + Length(w) + 1|w +
  - LL_base = Reading Time ~ Frequency(w) + Length(w) + 1|w + ...
- Delta of the log-likelihood tells us how much surprisal improved model fit
  - delta_ll = LL_base - LL_surp

Larger is (sometimes) better (in LMs)?

- Wilcox et al. 2020: positive correlation of reading time fit and perplexity

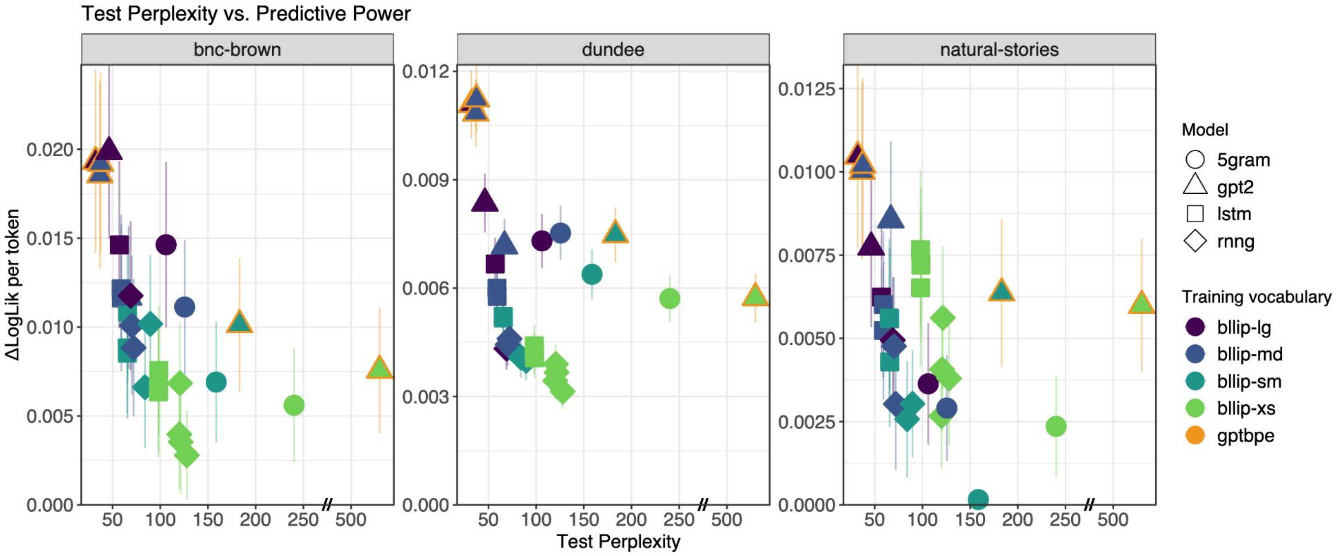
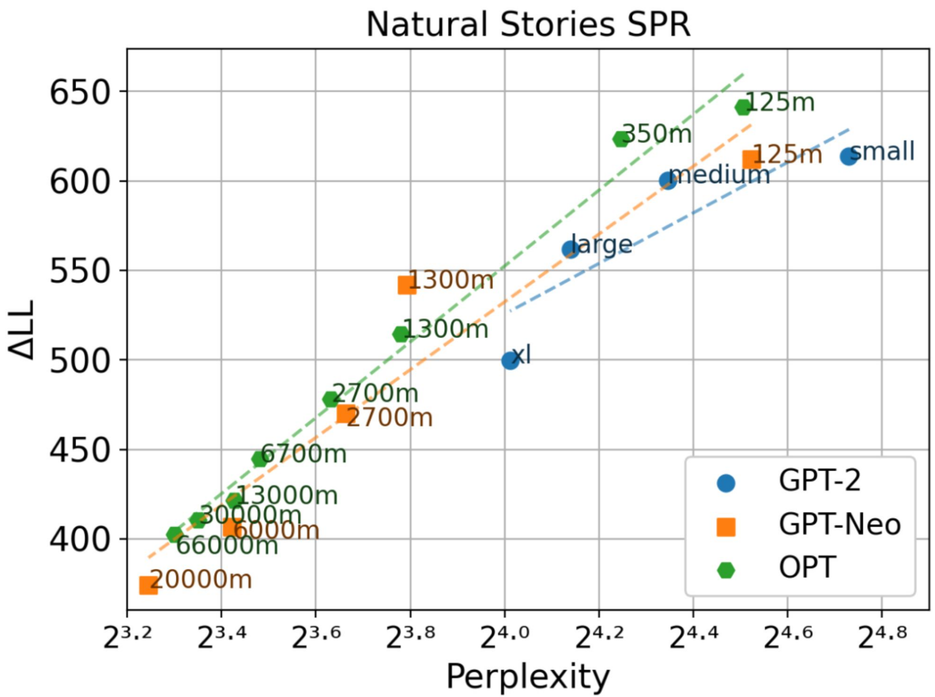


Figure 2: Relationship between predictive power (ΔLogLik) and model perplexity. Error bars are standard errors of by-fold mean ΔLogLik per token, using 10-fold cross validation. As model perplexity decreases, predictive power increases for all test corpora.

Larger is not always better (in LMs)!

- Oh & Schuler 2023
- LME(Reading Time ~ Surprisal)
- Inverse scaling of perplexity (model size) & reading time fit
- Oh & Schuler 2024: larger models memorize infrequent items -> underestimate surprisal



Jupyter Notebook