Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?

Byung-Doh Oh, William Schuler (TACL2023)

紹介者：栗林樹生（MBZUAI）

人間の言語獲得・処理が知りたい（認知科学からの要請）

難しさ

- 人間が人間（自分）について内省してしまうと科学の客観性が失われる
- しばしば直接な仮説の検証ができない
  - 頭を開いて脳を直接観察しても文法は書いてない
  - 子供を2グループに分けて統制して育てる…（倫理的に困難）

妥協点

- 構成論的アプローチ
  - 人間と同じ振る舞いをするものを作る．その作り方から可能な仮説を提示する．
  - その仮説が人間に対する説明として妥当かは別のレイヤから議論が必要

人間の言語処理が知りたい（認知科学からの要請）

- 計算理論のレベル
  - 何を計算しているか？計算の目的はなにか？（目的関数）
- データ構造・計算方法のレベル
  - どのように計算しているか？（モデルアーキテクチャ，内部表現）
- 物理的実装のレベル
  - （脳で）どのように実現・実装されているか？（ハードウェアレベルの実装）

[Marr, 1982]

人間が文を読んでいるときに何を計算しているか？（計算理論のレベル）

- サプライザル理論 [Levy08, Simth&Levy13, Shain+23]
  - 人間は文を読むときに先の単語を予測しており，予測が外れると処理に負荷がかかる
  - 各単語w_iの処理負荷は$-\log p(w_i|\boldsymbol{w}_{<i})$に比例する

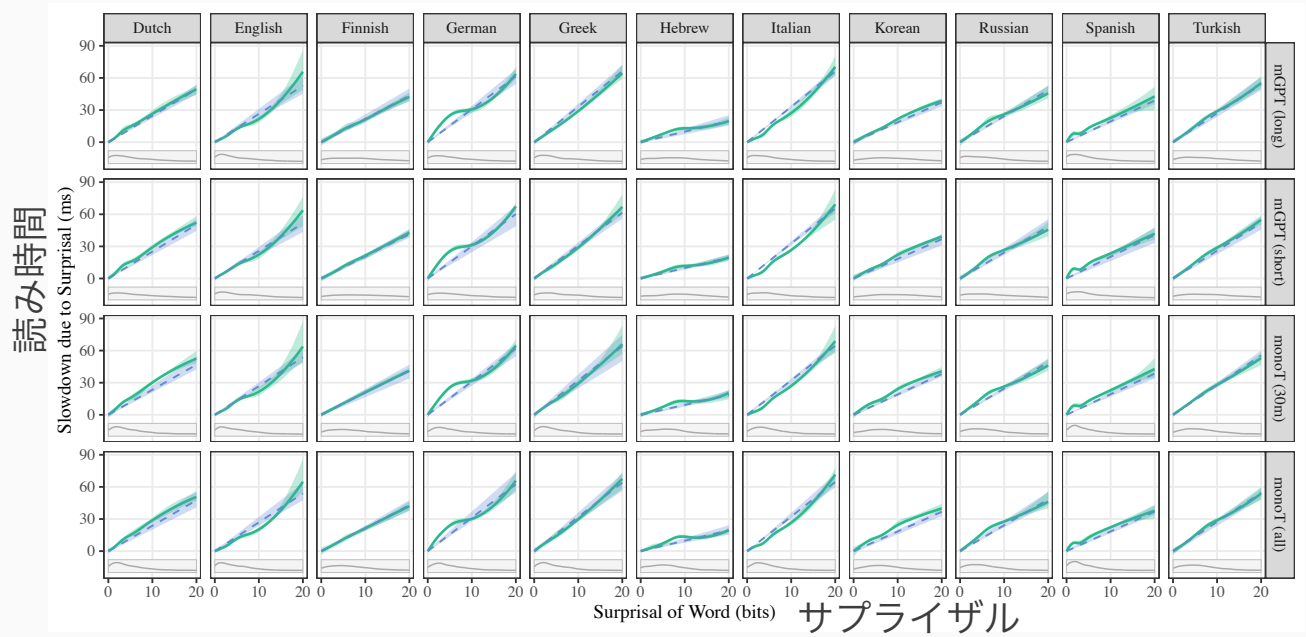


Figure 5: **Surprisal vs. Reading Time Relationship:** Non-linear GAMs are in green; linear control GAMs are in dotted blue. Shaded regions represent bootstrapped 95% confidence intervals. Grey subplots indicate the distribution of surprisal values. We find that GAMs recover a linear relationship between surprisal and reading-time slowdown.

直線的な関係で良いのかという点は，
昨年の勉強会資料も参照

https://speakerdeck.com/kuribayashi4/zui-xian-duan-nlplun-wen-shao-jie-revisiting-the-uniform-information-density-hypothesis-emnlp2021-linguistic-dependencies-and-statistical-dependence-emnlp2021

近年，エントロピー（サプライザルの期待値）の説明力の高さも改めて報告されているが，広く見ればこれもサプライザル理論だと括られている

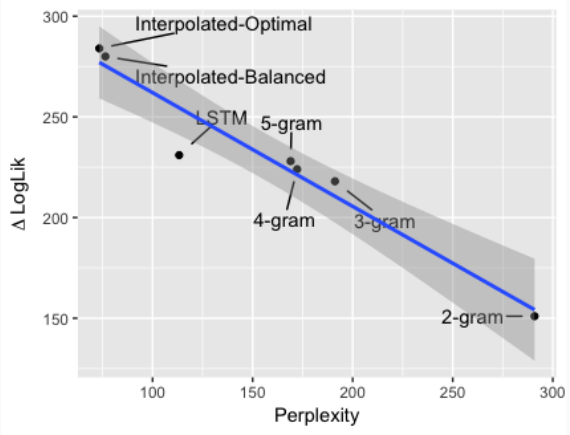
On the Effect of Anticipation on Reading Times (Pimentel+23)

Testing the Predictions of Surprisal Theory in 11 Languages (Wilcox+23)

どのようなモデルでサプライザルを計算するとよいか？（計算方法のレベル）

- データ構造・計算方法のレベル
  - どのように計算しているか？（モデルアーキテクチャ，内部表現）
  - どの程度正確な言語モデルで人間の文処理を説明できる？

言語モデルの性能が向上すると人間の振る舞いの説明もうまくできると経験的に信じられていた（~2022）



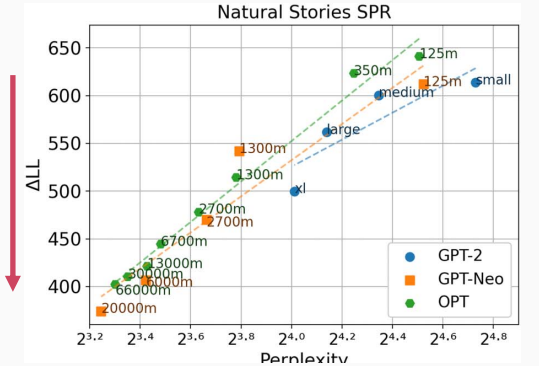
Predictive power of word surprisal for reading times is a linear function of language model quality (Goodkind+,18)

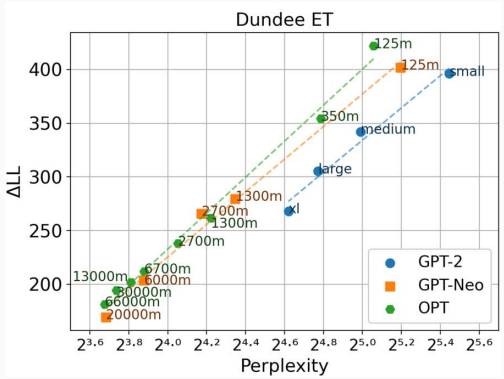
salve et al., 2012). It has even been observed that a language model's perplexity[4] correlates negatively with the psychometric predictive power provided by its surprisal estimates (Frank and Bod, 2011; Goodkind and Bicknell, 2018; Wilcox et al., 2020). If these language models keep improving at their current fast pace (Radford et al., 2019; Brown et al., 2020), exciting new results in computational psycholinguistics may follow, connecting reading behavior to the statistics of natural language.

Analyzing Wrap-Up Effects through an Information-Theoretic Lens (Meister+,22)

言語モデルの性能向上で見えてきた，認知モデリングにおけるスケーリング則（モデルのPPL∝読み時間の説明力）の破綻

人間はそこまで正確な次単語の予測ができていなさそう（サプライザル計算方法がある種貧しい）





図は紹介論文より

~300M params. 言語依存で破綻

Lower perplexity is not always human-like (Kuribayashi+,21)

we re-examine an established generalization —*the lower perplexity a language model has, the more human-like the language model is*— in Japanese with typologically different structures from English. Our experiments demonstrate that this established generalization exhibits a surprising lack of universality; namely, lower perplexity is not always human-like.

~1.5B params. 日英で破綻

Context limitations Make Neural Language Models More Human-Like (Kuribayashi+,22)

Notably, we also observed that **larger** GPT-2s have **less** human-like behavior in the full setting (right-most column in Table 4). This trend was weakened by introducing our context limitation.

~175B 英語で破綻

Large-Scale Evidence for Logarithmic Effects of Word Predictability on Reading Time (Shain+,23)

Our results additionally differentiate computational models of human next-word prediction. Surprisal estimates from GPT-2(-small) (Radford et al., 2019) substantially outperform surprisal estimates from *n*-gram, PCFG, GPT-J, and GPT-3 models. GPT-2 therefore appears to reside in a "Goldilocks" region of psycho-

~4.5B 13言語で破綻

Scaling in Cognitive Modelling: a Multilingual Approach to Human Reading Times (Varda+,23)

integration. This result corroborates the previous claims that cognitive modelling might constitute an exception to empirical scaling laws in NLP (Oh and Schuler, 2022). However, predictability es-

~13B? 11言語で破綻

Testing the Predictions of Surprisal Theory in 11 Languages (Wilcox+23)

但し同一言語で複数のモデルを比較しているわけではない

et al., 2020). However, studies on Japanese have failed to replicate these results, suggesting that the relationship does not hold for all languages (Kuribayashi et al., 2021). Further, Oh and Schuler (2023) and Shain et al. (2022) show that this relationship may not hold even in English for the most recent language models. To investigate this,

本研究：なぜ言語モデルの計算する確率は人間の読み振る舞いから逸脱していく？

- 基本方針：スケーリングが破綻しているコーパスの部分集合を見つけ，その言語的性質を眺める
  - 線形モデル：reading time ~ surprisal + baseline features (word frequency, word length…)
  - 特定の言語特性（例えば特定のPOS）をもつ部分コーパスごとにMSEを観察
  - サプライザルを計算する言語モデルを変えて同様に線形モデルを訓練し，どの部分コーパスでスケーリング則（モデルのPPLとMSEの正の相関）が破綻しているかを調べる

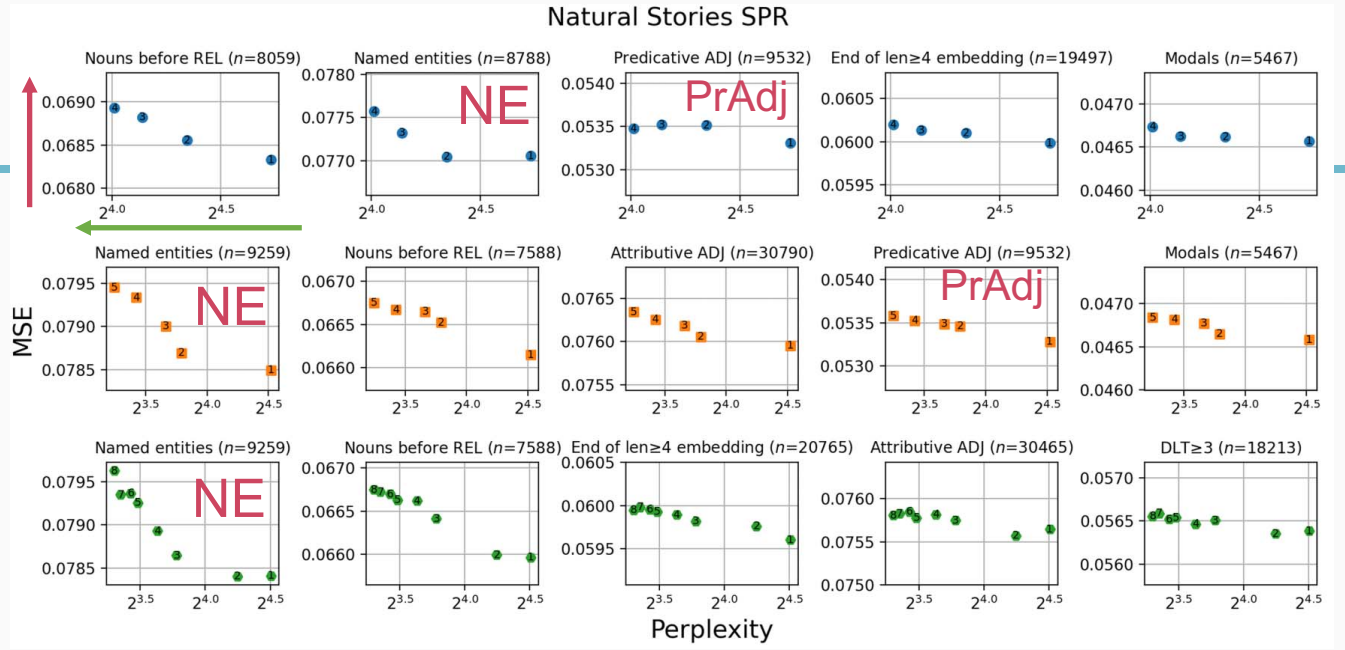
図は紹介論文より

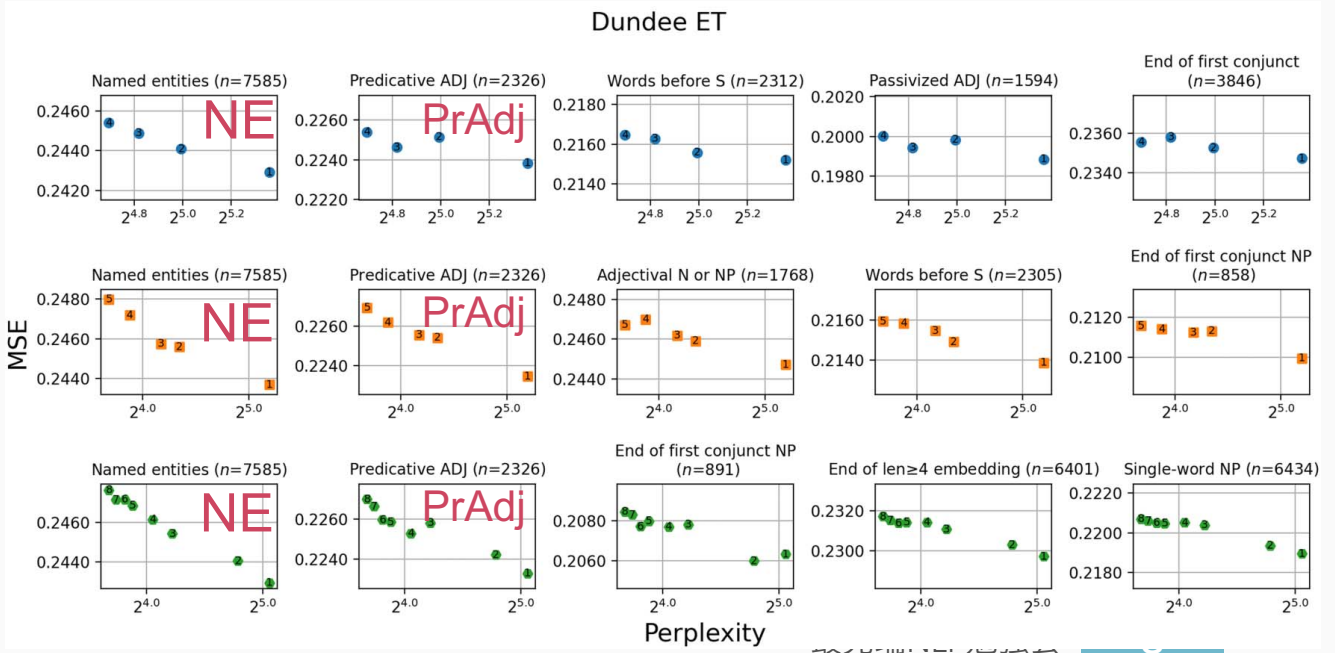
主な分析結果

- モデル×コーパス横断的に特定データポイントでスケーリングが破綻
  - Named Entity
  - Predicative ADJ
  - Nouns before REL (e.g., that)

私がもう少し包括的な説明をしにいこうと考え中…

- 基本的にモデルが読み負荷を過小に推定している
  - 人間に比べて「驚き」が小さすぎる
  - これを示す図は省略
  - 言語モデルが驚かなさすぎる観察は統語分析でもあり [Wilcox+,21]





感想

- ややポジションペーパっぽい
  - この文脈でスケーリングが破綻することを明言してくれた
- 知見は非常に観察的
  - オープンクラス（名詞，動詞，形容詞等）の語で言語モデルのサプライザルが低すぎる以上の一般化はなされていない
  - この分野において，初手一番大きいモデルを試すべきではないという教訓にはなる
- 認知的・言語学的な解釈とその裏付けが今後の課題
  - 人間の言語処理の効率（good-enough processing）や予測だけでは説明のつかないある種の制約（e.g., memory access, lexical access）の説明に繋がると期待
- スケーリングで解けない問題かつ，「次単語予測に関する科学」の一方向として，認知モデリングにさらに注目が集まることを期待
  - 最近TACLでアクティブな印象