Predictive power of word surprisal for reading times is a linear function of language model quality

Adam Goodkind & Klinton Bicknell
Northwestern University
Cognitive Modeling & Computational Linguistics Workshop

PROBABILITY IN CONTEXT

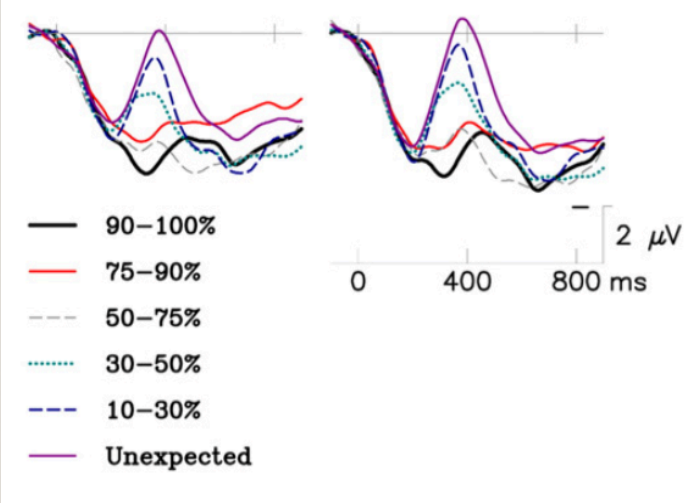
- *Don't touch the wet _____*
  - *paint* 
  - *cement* 
  - *bed* 

(Wlotko & Federmeier, 2012)

MOTIVATION
HOW WE USE PROBABILITY IN CONTEXT

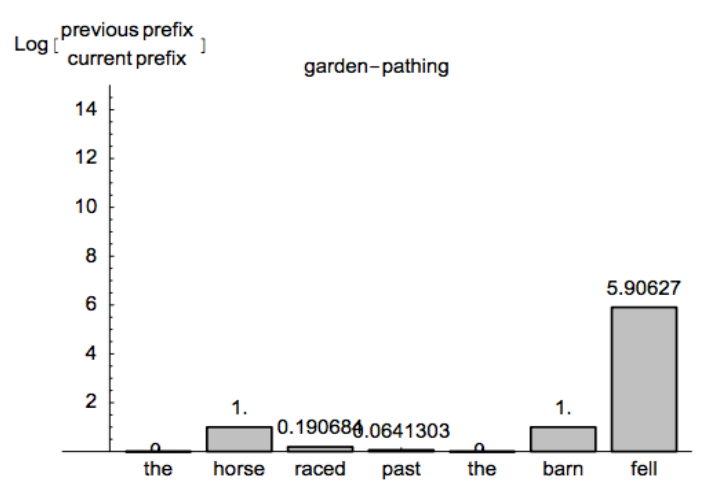
- Studies of human sentence processing have shown that a word's probability in context is strongly related to processing difficulty
- Do better estimates of word probability improve processing predictions?

ERP Response



(Wlotko & Federmeier, 2012)

Reading Times



(Hale, 2001)

SURPRISAL AND SURPRISAL THEORY

- From information theory (Shannon, 1948)
  - A theory of communication
  - The **information content** in a word = $-\log(p)$
- More information is more difficult to process
- Difficulty (cognitive cost of processing a word) ≈ how predictable the word is in a given context

$$difficulty \propto -\log P(w_i|w_{1\ldots i-1}, \text{CONTEXT})$$

(Hale, 2001; Levy, 2008)

- Prior studies (e.g. Demberg & Keller, 2008) found that surprisal can predict reading times

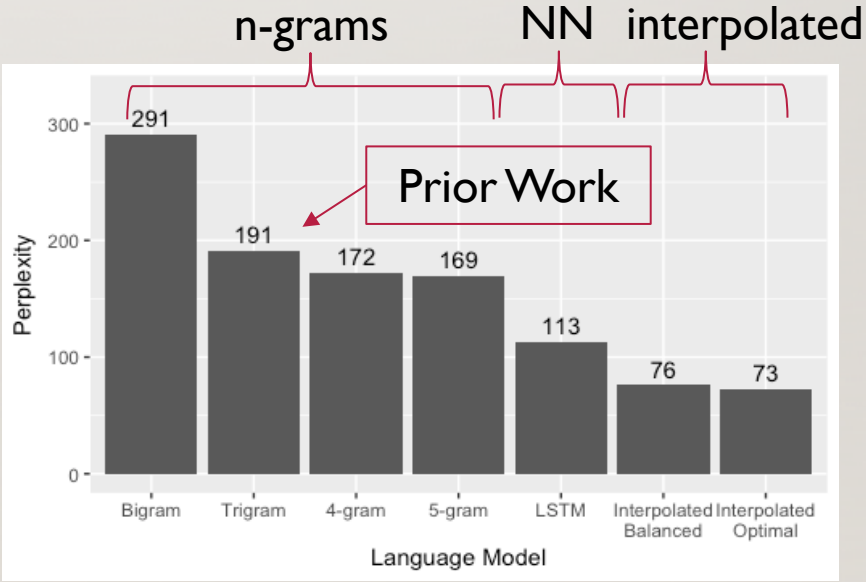
LANGUAGE MODELS
CALCULATING WORD PROBABILITIES

- Cloze task (Taylor, 1953)
  - Count people's responses to filling in a missing word
  - Inaccurate and labor intensive → need for computational models
- Language models
  - A probability distribution over sequences of words
  - Good language models assign a higher probability to word strings that occur more often
  - Quality (accuracy) of a language model is quantified as **perplexity**
    - Lower == Better

MANY TYPES OF LANGUAGE MODELS
DIFFERENT BUILDING BLOCKS

- *n*-grams (fixed sequence length)
  - Bigrams, trigrams, 4-grams, etc.
  - $p(w_n \mid w_{n-1})$
  - Fixed dependency length
- Neural network
  - Word probabilities use dependencies spanning arbitrary distances (number of words)
  - Usually use Long Short-Term Memory (LSTM) networks
  - Variable dependency length
- Interpolated
  - Combine multiple models

- Recent neural network-based language models have significantly improved linguistic accuracy



DEFINING "ACCURACY"

- `Linguistic accuracy`
  - How well language models predict unseen language
  - Measured by perplexity
- `Psychological accuracy`
  - How well language models predict psychological phenomena
    - E.g. eye gaze duration, ERP response amplitude

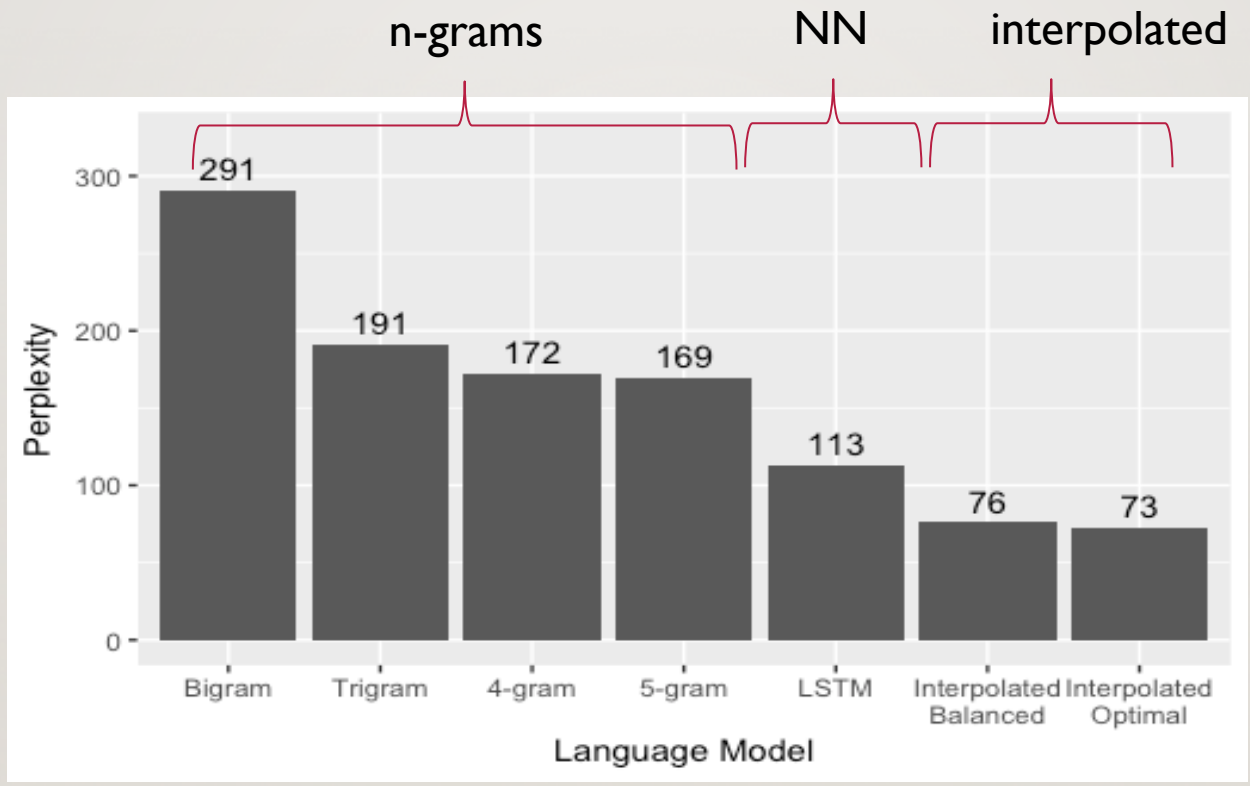
OUR STUDY

- Build a range of different types of language models
  - Different language models produce different estimates of surprisal
- Construct a regression model predicting gaze duration in an eye-tracking corpus from the surprisal of each language model
- Compare the regression models' quality of predictions for the gaze durations
  - Understand the relationship between language model quality and predictions of processing difficulty

METHODS
CREATING A LANGUAGE MODEL

- Language models used Google One Billion Word Benchmark ("1b") Corpus
  - Collected from international English news services
  - ~900 million words, 800,000 word vocabulary size
- n-grams models created with `kenlm`
  - Kneser-Ney smoothing
- Neural network model created from Google's pre-trained models
  - Long Short-Term Memory (LSTM) units in a Recurrent Neural Network (RNN)
- Interpolated models created by mixing LSTM and 5-gram estimates

OUR LANGUAGE MODELS

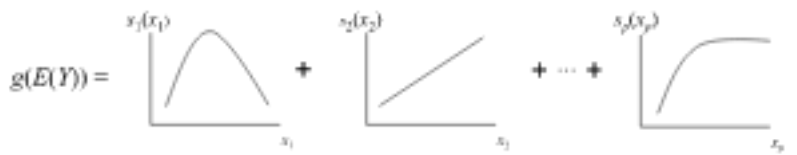


METHODS
EYE-TRACKING DATA

- Dundee Corpus
  - 61,000 tokens from a British newspaper, read by 10 participants
  - ~300,000 total tokens, 37,000 word vocabulary size
- Extracted gaze durations: how long a word was fixated during first pass reading
- Exclusions
  - Words not fixated
  - Words at beginning/end of line
  - …and others

METHODS
PREDICTIVE REGRESSION MODELS

- Generalized Additive Models (GAMs)
  - Type of regression model
  - Allows for non-linear effects



- Predictors of interest
  - Surprisal of current and previous words

METHODS
PREDICTIVE REGRESSION MODELS

- We used Generalized Additive *Mixed* Models (GAMMs)
- Predict eye gaze duration given:
  - **Surprisal of current and previous word**
  - Non-linear effects of control covariates
    - The interaction of word frequency and length
    - Sequential word number
    - Whether the prior word was fixated
    - Random intercepts for each subject

METHODS
PREDICTIVE REGRESSION MODELS

- Linear versus non-linear GAMMs
  - First set of experiments forced surprisal to be a linear predictor
  - Second set of experiments allowed surprisal to make non-linear predictions
    - Other predictors remained non-linear

METHODS
PSYCHOLOGICAL ACCURACY

- Measured improvements in predictions from each language model

$$\Delta\text{LogLik}(\text{model}_m) = \text{LogLik}(\text{model}_m) - \text{LogLik}(\text{baseline\_model})$$

  - LogLik (Log Likelihood)
    - A measure of accuracy
  - model_m
    - Includes language model *m*'s surprisal as a predictor
  - baseline_model
    - Missing predictor of interest (surprisal)
    - Includes only control covariates

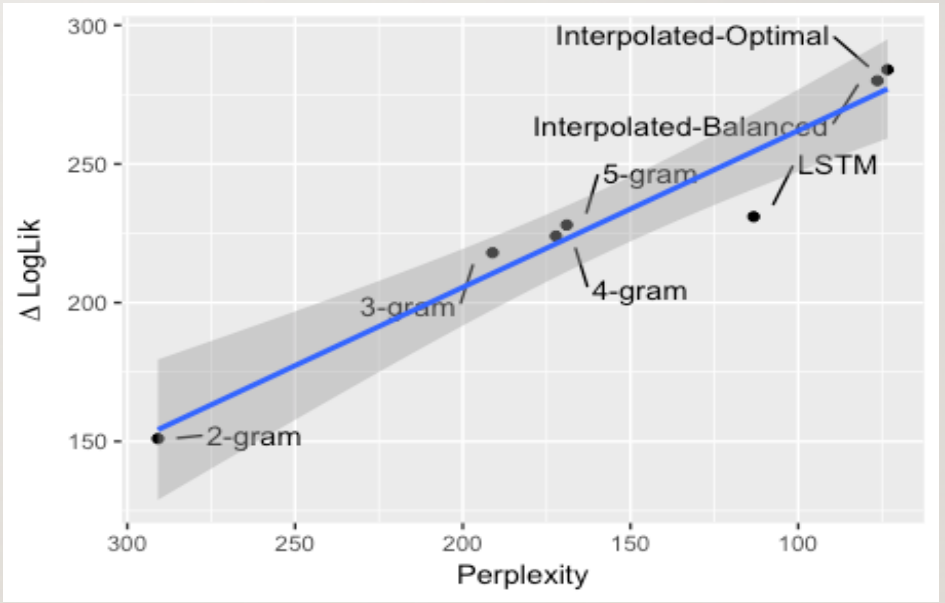
RESULTS
RELATIONSHIP BETWEEN LINGUISTIC AND PSYCHOLOGICAL ACCURACY

- Using a linear regression model, we investigate the relationship between language models and their psychological predictions
- What is the relationship between linguistic accuracy (perplexity) and psychological prediction quality (ΔLogLik)?

RESULTS
RELATIONSHIP BETWEEN LINGUISTIC AND PSYCHOLOGICAL ACCURACY

- As the perplexity of a language model improves, the model makes more accurate predictions for reading times
- This relationship holds across model types

Linear GAMMs



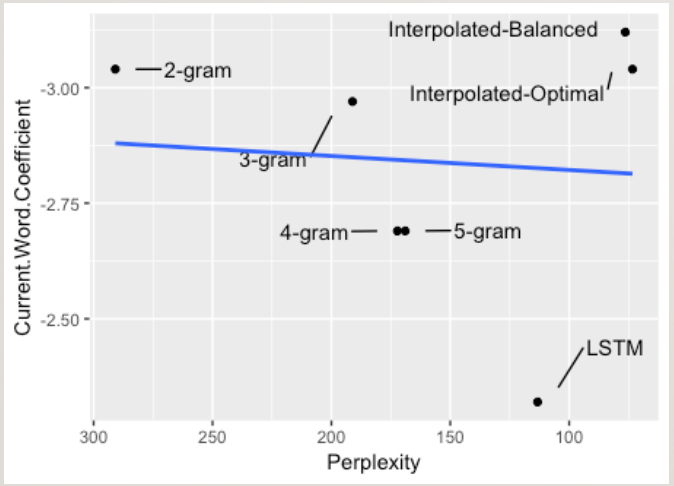
RESULTS
MAGNITUDE OF EFFECT

- As language models continue to improve and make better predictions, does the magnitude (size of effect) of surprisal change?
- Do better language models put more weight on the surprisal of current and previous words?
- We can compare coefficients of surprisal from each model to understand the magnitude of the effect

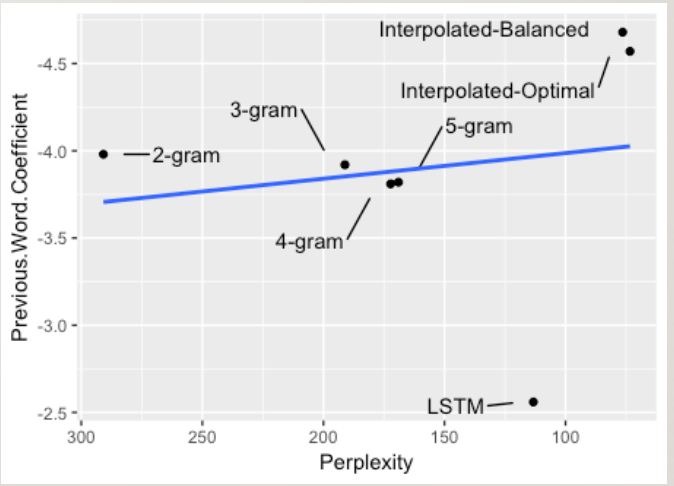
RESULTS
MAGNITUDE OF EFFECT

- The magnitude of the effect does not correlate with linguistic accuracy
- Effect size of surprisal does not seem to be biased for worse language models

Current word



Previous word



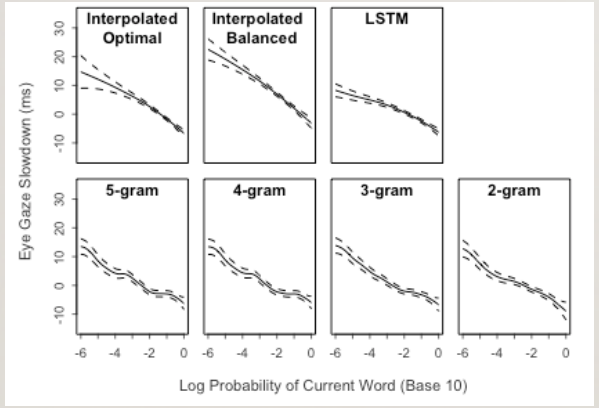
RESULTS
SHAPE OF EFFECT

- Smith & Levy (2013) looked at the shape of the effect of surprisal
  - Found a linear relationship
  - Supports various derivations of surprisal theory
    (e.g., Hale, 2001; Levy, 2008; Bicknell & Levy, 2009; Smith & Levy, 2013)
  - Contra alternative probabilistic processing theories
    (e.g., Narayanan & Jurafsky, 2004; theories predicting UID optimality)
- Does this linear relationship hold for more sophisticated models, if we allow surprisal to be non-linear?

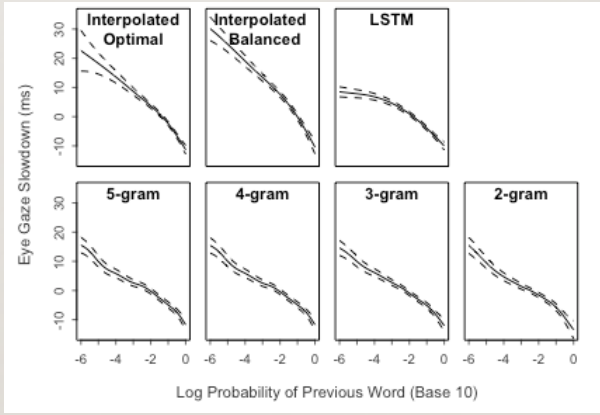
RESULTS
SHAPE OF EFFECT

- For both the current and previous word probability, gaze time changes at a linear rate, for all models
- Possibly even *more* linear as language model accuracy improves

Current word



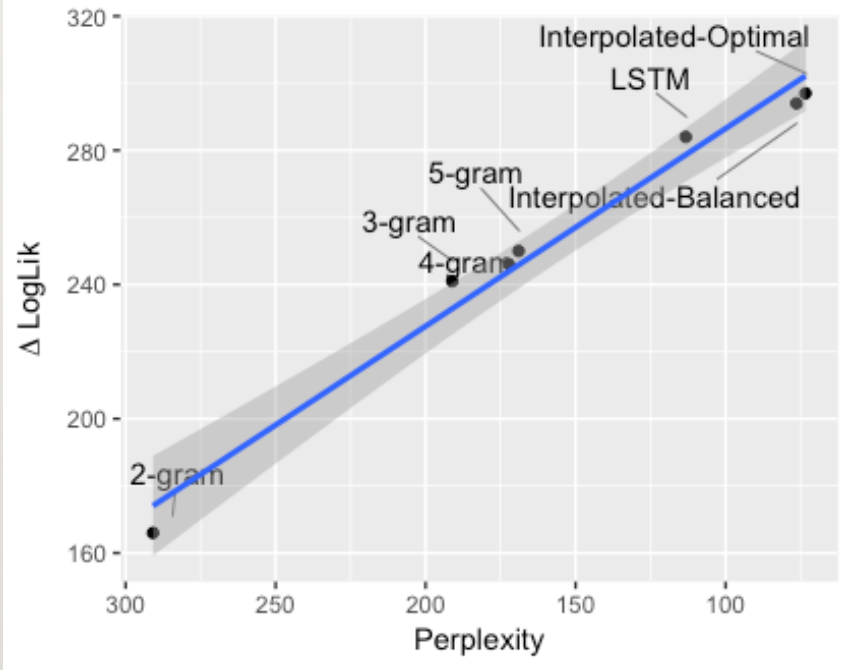
Previous word



RESULTS
RELATIONSHIP BETWEEN LINGUISTIC AND PSYCHOLOGICAL ACCURACY (PART II)

- If we allow for non-linear effects, not only does the relationship between models improve, but the relationship becomes *more* linear

Non-linear GAMMs



TAKEAWAYS

- Strong relationship between linguistic model quality and its psychological predictive power
  - No privileged language model class: better perplexity improves psychological predictions
- The size of the surprisal effect was consistent across models
  - Estimates of the effect size of surprisal from worse language models appear to be relatively unbiased
- The effect of surprisal is linear across all models and distributions of word probabilities
  - Supports surprisal theory processing models even with state-of-the-art language models

THANK YOU!

Funding sources:

Northwestern | Northwestern Institute on Complex Systems DATA SCIENCE

Adam Goodkind
a.goodkind@u.northwestern.eduz