The Ohio State University

Surprisal Estimators for Human Reading Times Need Character Models

Byung-Doh Oh, Christian Clark, William Schuler

Introduction

- Popular use of character models in NLP (Kim et al., 2016; Lee et al., 2017)
- Evaluation of surprisal estimates from word-level neural LMs (Goodkind & Bicknell, 2018; Futrell et al., 2019; Wilcox et al., 2020; Hao et al., 2020)
- Do character models give more predictive surprisal estimates?
  - Character model within a structural parser-based model
  - Comparison of predictive power on three different datasets

Surprisal: $-\log P(w_t | w_1 \dots w_{t-1})$

- Predictive of measures of processing difficulty (Hale, 2001; Levy, 2008)

- Left-corner parsers reflect memory and processing constraints (Miller & Isard, 1964; Johnson-Laird, 1983)
  - Limits on center embedding
  - Fixed number of operations at every word

Left-corner parsing

$$P(w_t\ q_t | q_{t-1})$$

$$= \sum_{l_t, g_t} P(l_t | q_{t-1}) * P(w_t | q_{t-1} l_t) * P(g_t | q_{t-1} l_t w_t) * P(q_t | q_{t-1} l_t w_t g_t)$$

- Hidden states q_t consist of derivation fragments
- Marginalize over hidden states q_t for prefix probabilities

Left-corner parsing

$$P(w_t\ q_t|q_{t-1})$$

$$= \sum_{l_t, g_t} P(l_t|q_{t-1}) * P(w_t|q_{t-1}l_t) * P(g_t|q_{t-1}l_tw_t) * P(q_t|q_{t-1}l_tw_tg_t)$$

- Hidden states q_t consist of derivation fragments

- Marginalize over hidden states q_t for prefix probabilities

Left-corner parsing: w_t

$$P(w_t|q_{t-1}l_t) = \sum_{x_t,r_t} P(x_t|q_{t-1}l_t) * P(r_t|q_{t-1}l_t x_t) * P(w_t|q_{t-1}l_t x_t r_t)$$

- To a lemma x_t, apply a morphological rule r_t for word w_t
- Morphological rules come from a GCG annotation scheme (Nguyen et al., 2012)
- Character-based RNN sub-models for estimating $P(x_t|q_{t-1}l_t)$ and $P(r_t|q_{t-1}l_t x_t)$

Surprisal estimation

- Parser trained on GCG reannotation of WSJ02-21 (Marcus et al. 1993)

- Surprisal estimated using beam search
  - Test: Character-based word generation model (*CharWSurp*)
  - Baseline: Relative frequency estimation (*FreqWSurp*)

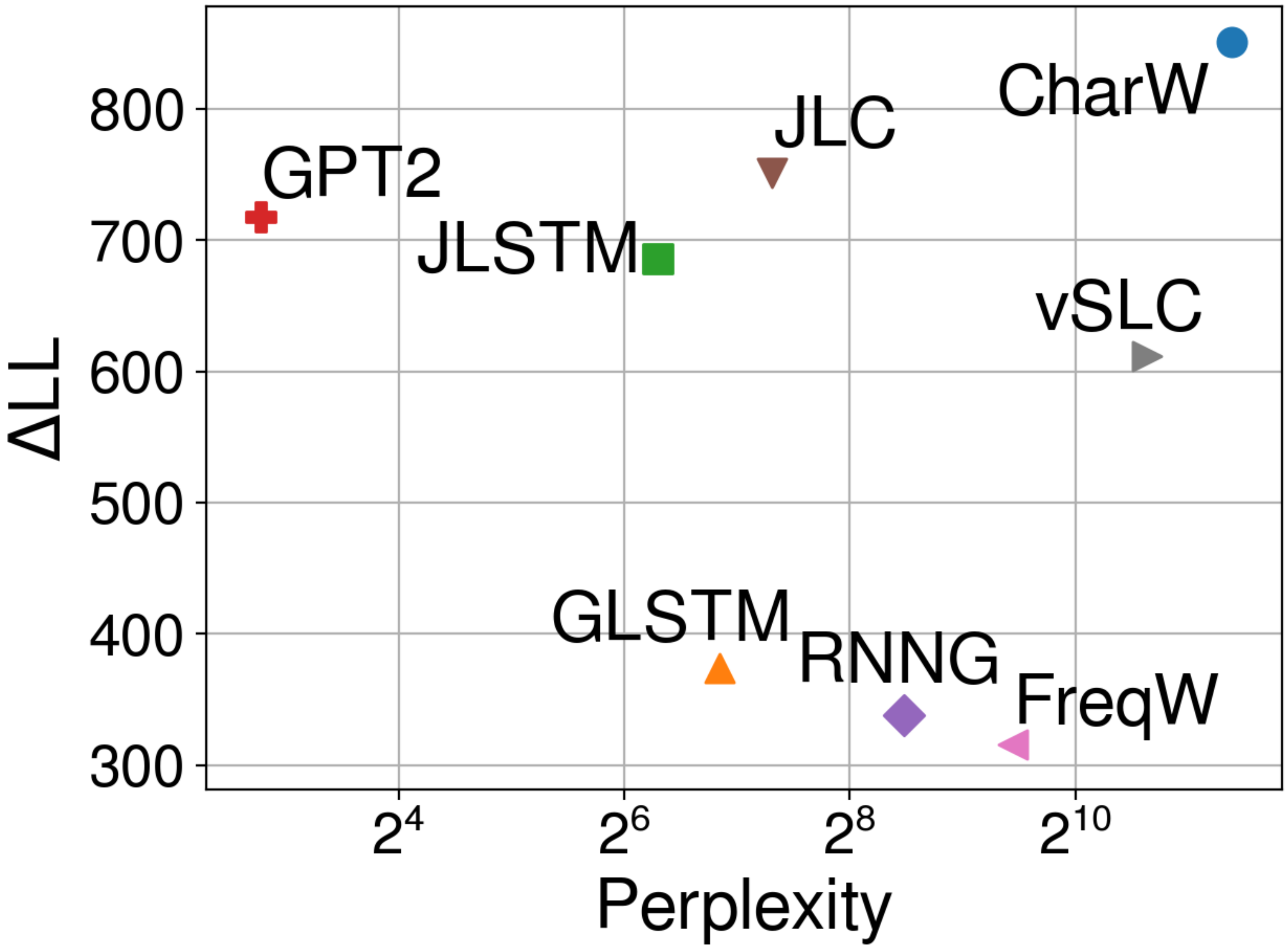
Psycholinguistic evaluation

- Comparison against surprisal estimates from various models
  - GLSTM (Gulordava et al., 2018), JLSTM (Jozefowicz et al., 2016), GPT2 (Radford et al., 2019)
  - RNNG (Hale et al., 2018), vSLC (van Schijndel et al., 2013), JLC (Jin & Schuler, 2020)

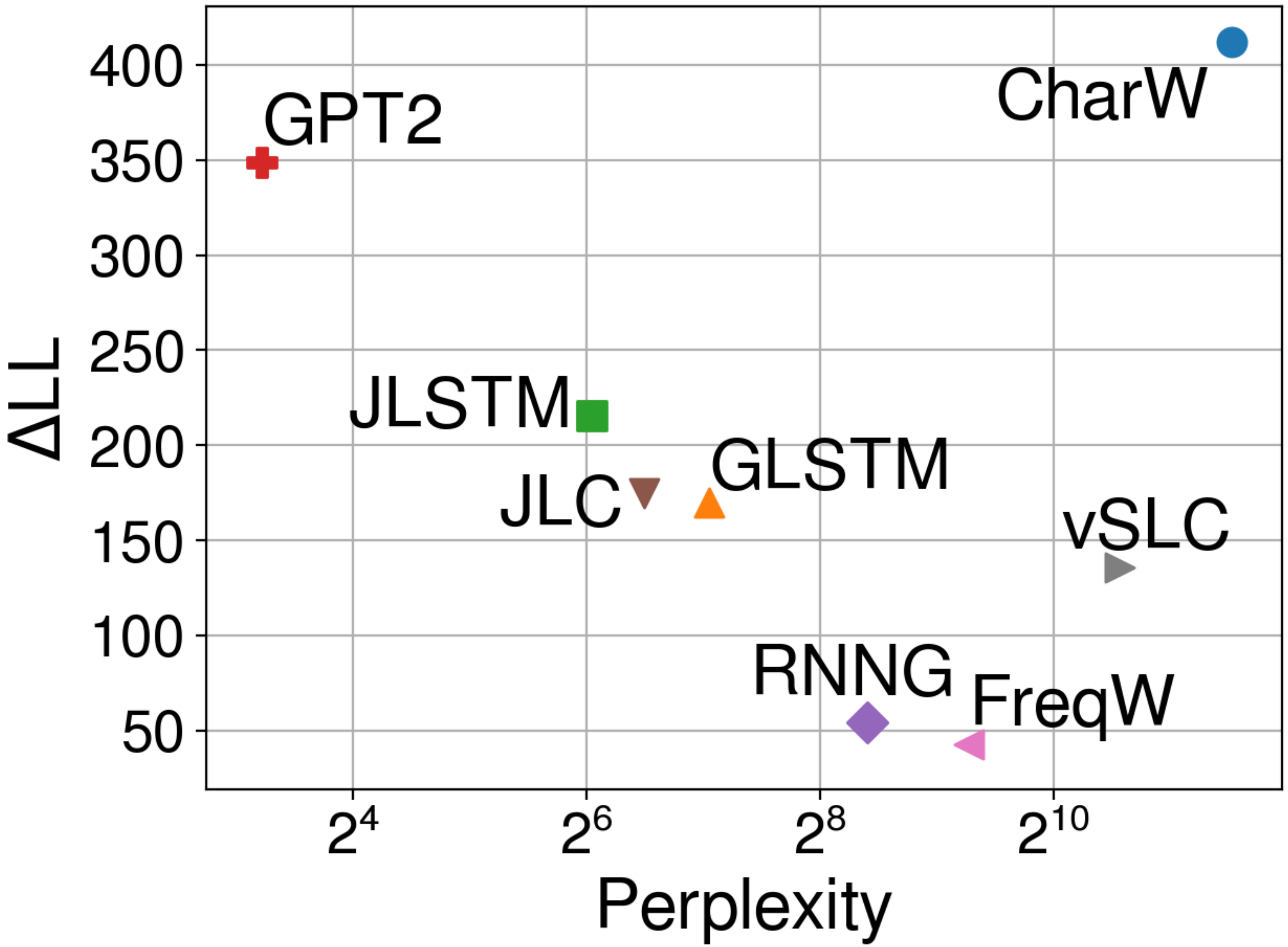
- Evaluation metric: Δlog-likelihood (Goodkind & Bicknell, 2018; Hao et al., 2020)

- Evaluation on three datasets
  - Natural Stories SPR (Futrell et al., 2018), Dundee ET (Kennedy et al., 2003), Natural Stories fMRI (Shain et al., 2019)

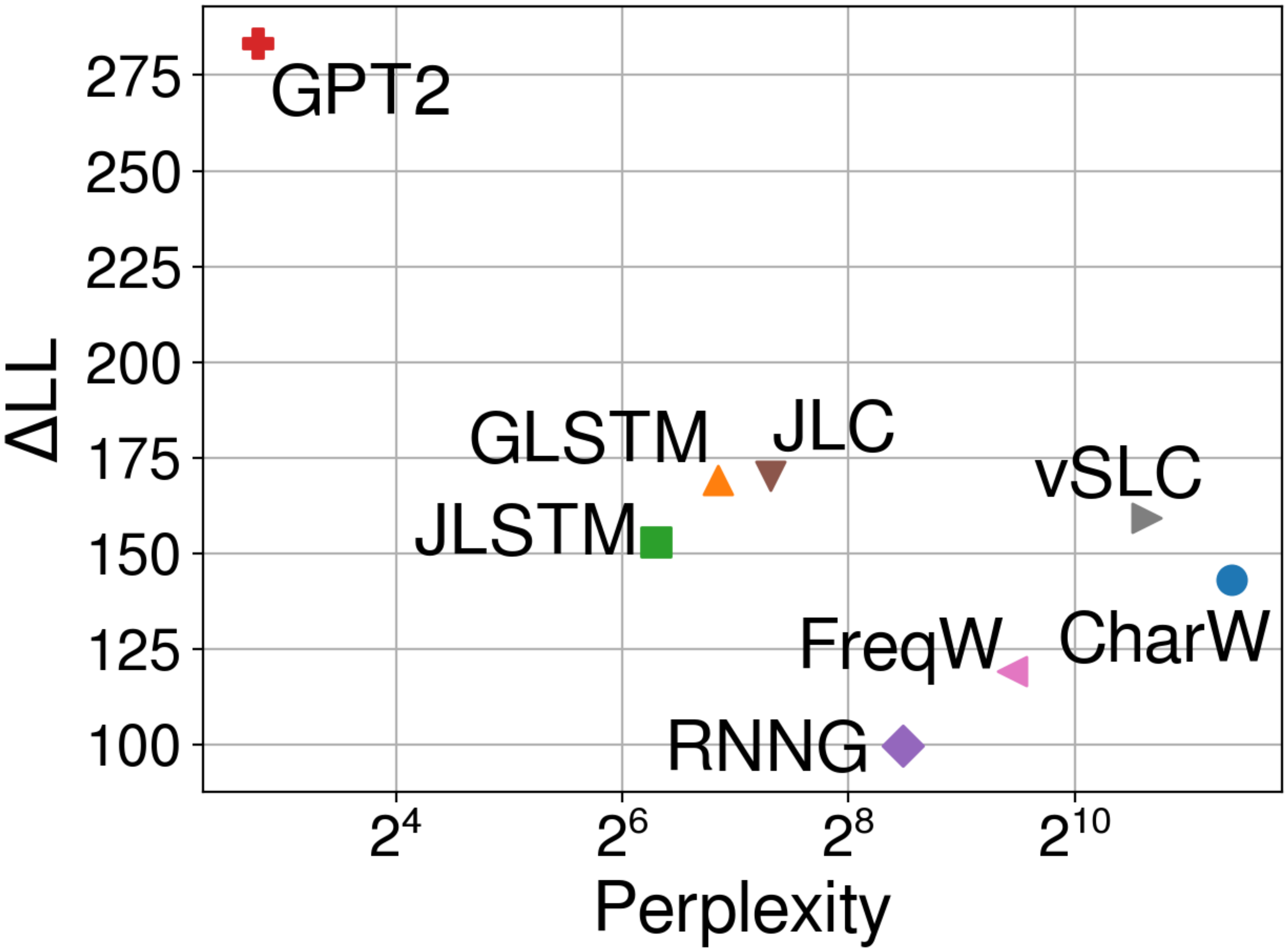
Results (Natural Stories SPR)



Results (Dundee ET)



Results (Natural Stories fMRI)



Conclusion

- Character model for word generation probabilities within a parser
- Contributes to better fits to human response data
  - Better than large-scale neural LMs on SPR and ET data
- New nuance to the relationship between PPL and predictive power
  (Goodkind & Bicknell, 2018; Wilcox et al., 2020)

THE OHIO STATE UNIVERSITY

Thank you!

This work was supported by NSF Grant #1816891.

Code for this work is publicly available at https://github.com/byungdoh/acl21_semproc.