The architecture gave the language model the constituent length preferences

Neil Rathi
Stanford University
rathi@stanford.edu

One of the hallmarks of modern language models is their ability to generate text that is not just grammatical but also *natural*. Futrell and Levy (2019), for example, show that RNN models are capable of recovering human-like 'soft' ordering preferences. Here, we show that some such preferences can arise from architectural considerations on memory, rather than purely from patterns in the training data. This has particular implications for models of *human* language production.

Consider the English **ditransitive alternation**:

(1) The teacher gave $\overbrace{\text{the student}}^{\text{RECIPIENT}}$ $\overbrace{\text{the book}}^{\text{THEME}}$.

(2) The teacher gave the book to the student.

Sentences (1) and (2) have the same semantic content, but differ in their relative order of recipient and theme.[1] This ordering is in large part determined by **availability effects**, where some constituents are more 'available' than others and are thus produced earlier (Levelt, 1981; Bock, 1982). In practice, this manifests in featural preferences: animate before inanimate, definite before indefinite, short constituents before long ones (Stallings and MacDonald, 2011; Koranda et al., 2022).

Futrell and Levy (2019) show that such featural preferences also emerge in an LSTM RNN trained on natural language data. Some of these effects very clearly boil down to the fact that human ordering preferences are reflected in human data—there is nothing inherent to the architecture of an LSTM that would induce a bias towards producing animate or definite nouns earlier if such a bias was not present in the training data.

Here, however, we show that one such preference (namely, constituent length) will emerge out of any language model with an imperfect memory representation, *even when trained on data that reflects no ordering preference*. We first demonstrate that a language model with a 'perfect' representation of memory will *not* induce any ordering effects. We then provide empirical evidence demonstrating that on a controlled dataset, statistical n-gram models and LSTMs with imperfect memory induce human-like constituent length preferences.

To make this more concrete, define a language model as a stochastic **policy** $\pi(u_n \mid s)$, a distribution on utterances u_n given context s. In a 'perfect' language model, s is exactly equal to $u_{1:n-1}$. We can describe a 'general case' dative alternation as

(3) The a VERB the c_1 (to) the c_2,

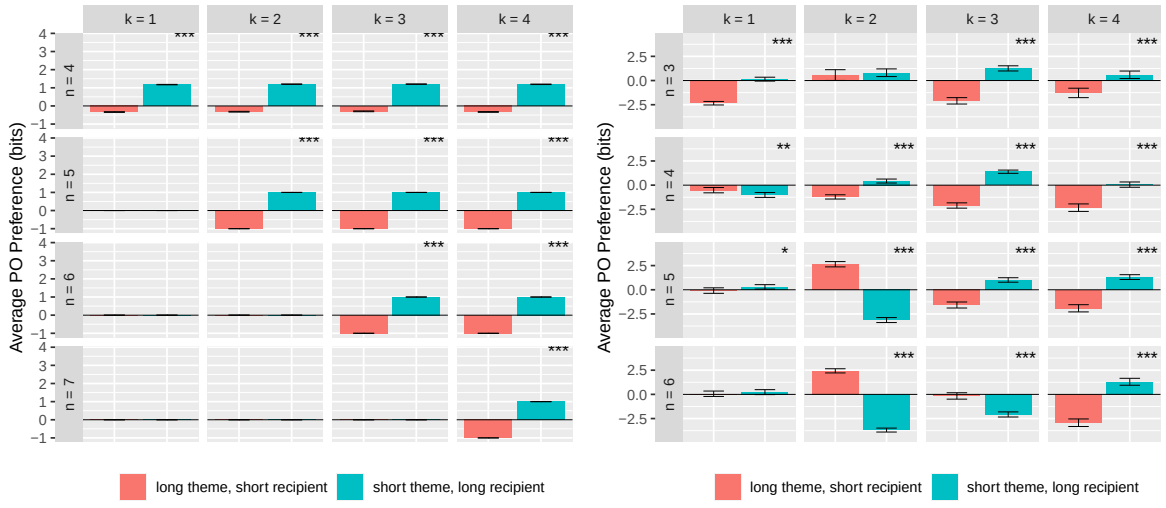
where each $c_1, c_2 \in \{t, r\}$ is either a theme or recipient. If we assume that our perfect language model has been trained on perfectly unbiased training data, we have $p(c_1 = t) = p(c_1 = r)$. But once the model produces c_1, it implicitly seals the fate of c_2. Any perfect model will thus have no preference for either order of t and r, regardless of constituent length, because initial selection of c_1 is always sufficient to determine c_2.

Now, however, suppose we have a *lossy* representation of memory where $s = M(u_1, \ldots u_n)$ is some encoded version of the actual context. Now, the model may not know if $c_1 = t$ or r, depending on their length; thus, the ambiguity returns and a short-before-long preference can emerge. The actual derivation of this depends on the nature of the loss function (see Futrell et al., 2020; Hahn et al., 2022), so we focus here on a naïve loss model: **fixed-length context windows**, where the language model only has access to the previous n tokens to make autoregressive predictions.

Here, we empirically evaluate two architectures with such fixed contexts: statistical n-gram models and fixed-context LSTMs, manipulating the size of their context window.

To avoid any data-driven effects regarding constituent length, we train these models on a simu-

[1] Sentences of the form of (1) are called 'double object' (DO) constructions, while sentences like (2) are 'prepositional object' (PO) constructions.



(a) n-gram (b) LSTM

Figure 1: Average preference for prepositional object order, by theme/recipient length, for n-gram and LSTM models with context window size n with 'long' constituent length $k + 1$. Error bars represent 95% confidence interval; plots marked with significance of t-test between conditions. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

lated dataset consisting solely of dative sentences. We first generate a set of tuples each consisting of two nouns and two 'adjective phrases,' uniformly drawn from a simulated vocabulary.

However, with a too-short adjective phrase and a too-long context window, the model is effectively equivalent to the perfect model described above. Thus, for each context window size, we also manipulate the length of the adjective phrase (between 1 and 4 tokens). In total, we train four models for every context window size.[2]

Then, for each tuple, we generate all sixteen permutations of sentence (3), where a consituent c is a concatenated adjective phrase and noun phrase. The training set consists of 64000 such sentences.

We generate the 8000-sentence test set similarly. Of particular interest is the preference for producing short constituents before long constituents. For a given item in the test set (i.e. a theme and recipient), we measure the 'PO preference' as the difference in **surprisal** $S(w_{1:n}) = -\log p(w_{1:n})$ between the DO construction and the PO construction. Thus, if a model obeys human-like ordering preferences, we expect a positive PO preference when the recipient is longer than the theme, and a negative PO preference when it is shorter, as this corresponds to a short-before-long order.

[2]Each LSTM was trained with the same hyperparameter set (excluding context window length): batch size 20, embedding dimension 200, and learning rate 20. Each model was trained for 15 passes through the training data, with early stopping.

Figure 1 shows results for both n-gram and LSTM models. When the context window length is not too long (i.e. when the representation of the context is not equivalent to the actual context), we notice human-like preferences emerge, with the 'long recipient, short theme' case having consistently higher average PO preference than the 'short recipient, long theme' case. We run a t-test comparing the average PO preference for each condition in each model, finding a significant difference by condition for all applicable models.

Taken together, these results reveal that language models with an imperfect memory representation—in the form of a fixed length context window—can learn human-like 'soft' ordering preferences, without any biases in data. This is of particular interest to incremental models of *human* language production. Most of these models choose to eschew a concrete grounding of availability preferences, instead grouping these into an empirically operationalized 'cost' function (e.g. RSA, see Degen et al., 2020; Degen, 2023).

This work suggests that availability factors can come out of a more generic mechanism (e.g. a lossy automatic policy), as posited by Futrell (2023). Indeed, the idea of lossy memory has seen wide success in the related field of online language *processing* (Futrell et al., 2020; Hahn et al., 2022). Future work should examine whether and how other factors of availability might come out of such a generic model.

References

J Kathryn Bock. 1982. Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological review*, 89(1):1.

Judith Degen. 2023. The rational speech act framework. *Annual Review of Linguistics*, 9:519–540.

Judith Degen, Robert D Hawkins, Caroline Graf, Elisa Kreiss, and Noah D Goodman. 2020. When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*, 127(4):591.

Richard Futrell. 2023. An information-theoretic account of availability effects in language production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Richard Futrell, Edward Gibson, and Roger P Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, 44(3):e12814.

Richard Futrell and Roger P. Levy. 2019. Do RNNs learn human-like abstract word order preferences? In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 50–59.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119.

Mark J Koranda, Martin Zettersten, and Maryellen C MacDonald. 2022. Good-enough production: Selecting easier words instead of more accurate ones. *Psychological Science*, 33(9):1440–1451.

Willem JM Levelt. 1981. The speaker's linearization problem. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 295(1077):305–315.

Lynne M Stallings and Maryellen C MacDonald. 2011. It's not just the "heavy np": relative phrase length modulates the production of heavy-np shift. *Journal of psycholinguistic research*, 40:177–187.