Flexibility in Language Production: Insights from Completion of Fragmentary Inputs

Peng Qian, Roger Levy (MIT)

Naturalistic human language production is predominantly left-to-right, but we can generate and predict language in much more flexible ways. Literate speakers use this ability regularly in text editing, which often involves changing part of a sentence while respecting the constraints imposed by the parts left unchanged. In utterance planning, speakers may commit to use a particular word or phrase, forcing them to navigate from a sentence's beginning to arrive at it successfully. Although this flexible language generation ability is intuitively likely to be closely related to the mechanisms of language comprehension and production studied in psycholinguistics, it has thus far received comparatively little attention in sentence processing research. Here we report initial steps in advancing our understanding of this capability, under the hypothesis that these mechanisms of constrained linguistic generation are scaled-up versions of the same simple computational "motifs" that allow robust processing for degraded inputs (Samuel 1981; Dilley & Pitt, 2010), and follow principles of noisy-channel probabilistic inference (Levy, 2008; Gibson et al., 2013; Keshev et al., 2020).

We focus on the empirical problem of completing fragmentary linguistic input: for example, given an incomplete sentence such as "____ *easy* ____ *problem* ____", native speakers can quickly come up with reasonable completions for the missing pieces, and can even handle more challenging inputs like "*Vineyards were found scattered throughout* ____ *visited grew any grapes*". To gain insight into the mechanisms underlying these abilities, we use a reverse-engineering approach, evaluating the quality of a theory by its qualitative and quantitative fit to human behavioral data. We formalize the task of generating completions B from fragments C as Bayesian computation of the posterior $P(B|C)$, assuming a generative model over the space of all possible linguistic utterances. As a concrete instantiation of the "motif hypothesis", we built a neurally-guided sampling-based inference algorithm, GibbsComplete, consisting of a masked language modelling motif (BERT; Devlin et al., 2018) as the proposal distribution $P(B_i|B_{\setminus i},C)$ and a next-word language modelling motif (GPT-2; Radford et al., 2019) as the scoring function $\varphi(B, C)$, inspired by Wang & Cho (2019). Neither of the computational motifs is optimized for solving the exact target sentence completion task, in contrast to an alternative "fine-tuning hypothesis" of specialized mechanisms for fragmentary input completion, which we implement by tuning pretrained language models (ILM, Donuhue et al., 2020; BART, Lewis et al., 2019; T5, Raffel et al., 2019) to directly predict the completions B conditioned on a neural encoding of the input fragments C.

Our Study 1 evaluates models' abilities to follow global syntactic context subject to the grammatical constraints, using 26 sets of targeted tests featuring structural reasoning. Here, GibbsComplete's performance is comparable to fine-tuned models despite no specific training for the task (Figs 1, 2). Study 2 quantitatively compares models' match to item-level patterns of fragment completion, using 120 stimuli of the form "____ w_1 ____ w_2 ___." where w_1 and w_2 are single words (40 each Noun–Noun, Adj–Adj, Adj–Noun). We use the syntactic category of the least common ancestor of w_1 and w_2 in parsed completions as a statistic for human–model comparison. Here, GibbsComplete outperforms all the fine-tuned models (Figures 3, 4). These results provide initial support for our "motif" hypothesis, and open the door to new future investigations of how linguistic knowledge can be flexibly deployed by the human mind.

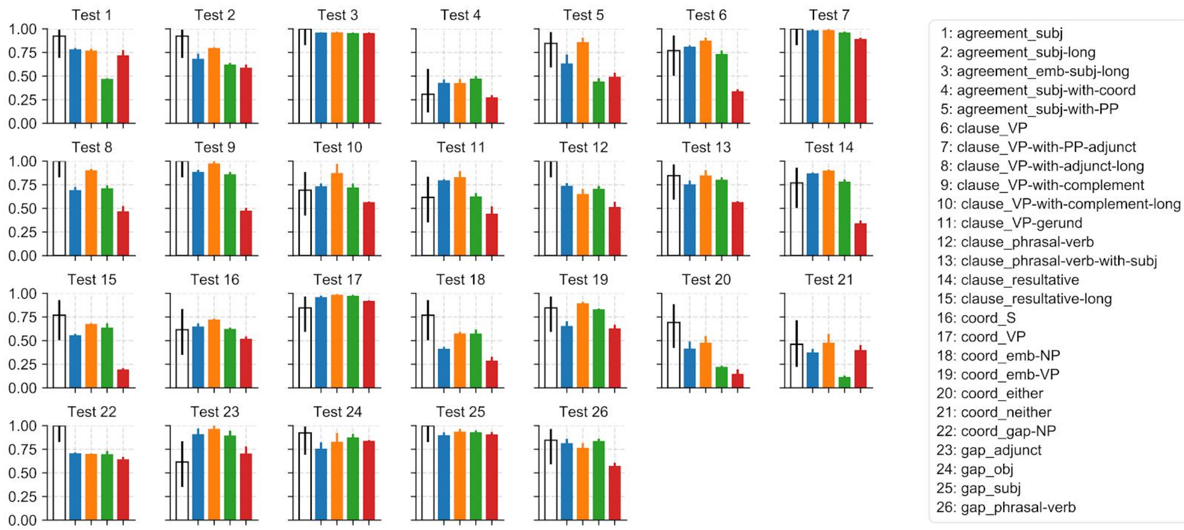

Figure 1: Models' performance in respecting grammatical constraints from fragments (Study 1)

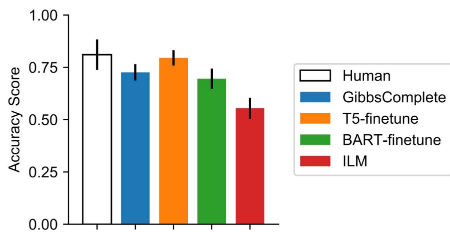

Figure 2: Aggregate performance (Study 1)

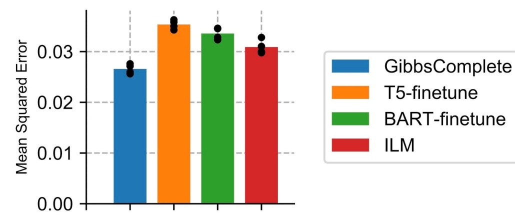

Figure 3: MSE to human completions (Study 2)

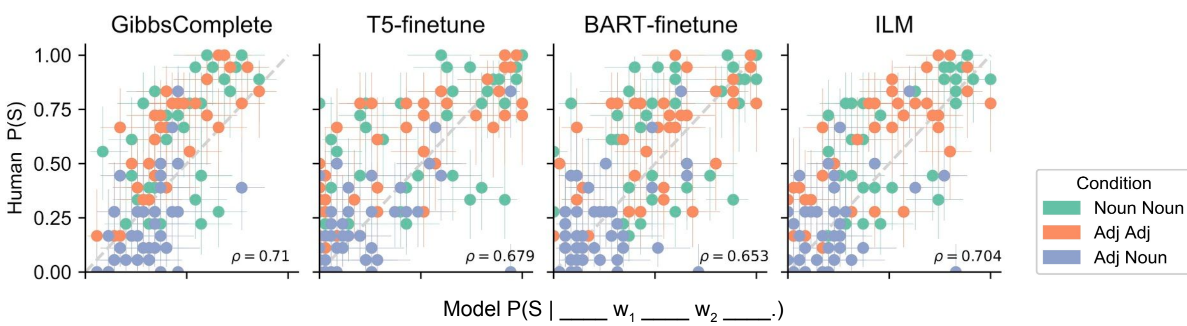

Figure 4: Comparing model output to human completions on the statistic of *S* as lowest common ancestor

References: Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL. • Dilley & Pitt (2010). Altering context speech rate can cause words to appear or disappear. Psychological Science. • Donahue, C., Lee, M., & Liang, P. (2020). Enabling Language Models to Fill in the Blanks. ACL. • Gibson et al. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. PNAS. • Keshev & Meltzer-Asscher. (2020). Noisy is better than rare: Comprehenders compromise subject-verb agreement to form more probable linguistic structures. Cognitive Psychology. • Levy (2008). A noisy-channel model of human sentence comprehension under uncertain input. EMNLP. • Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461. • Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog. • Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683. • Samuel (1981). Phonemic restoration: insights from a new methodology. JEP: General. • Wang, A., & Cho, K. (2019). BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. In Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation.