Masked language models directly encode linguistic uncertainty

Cassandra L. Jacobs
Department of Linguistics
University at Buffalo
cxjacobs@buffalo.edu

Ryan J. Hubbard & Kara D. Federmeier
Beckman Institute and Department of Psychology
University of Illinois at Urbana-Champaign
{rjhubba2;kfederme}@illinois.edu

1 Introduction

Recent advances in human language processing research have suggested that the predictive power of large language models (LLMs) can serve as cognitive models of human language processing. Evidence for this comes from LLMs' close fit to human psychophysical data, such as reaction times or brain responses in language comprehension experiments. Those adopting LLM architectures as models of human language processing frame the problem of language comprehension as prediction of the next linguistic event (Goodkind and Bicknell, 2018; Eisape et al., 2020), in particular focusing on lexical or syntactic surprisal. However, this approach fails to consider that comprehenders make predictions using some representation of the *content* of an utterance. That is, in contrast to surprisal, readers make use of a mental model that creates an evolving understanding of who is doing what to whom and how. In contrast to comprehenders, surprisal measures do not make predictions about the content, as surprisal simply measures the conditional probability of some linguistic event given the surrounding context.

Many convergent cues in the upstream context, such as the frequencies of words in a sentence so far, will affect hidden state representations of models, which may then influence the predictability of upcoming words. The present work deviates from the surprisal paradigm by assessing how much the hidden state representations of LLMs, which are the source of the predictive power that LLMs have over symbolic representations, encode human language processing-relevant uncertainty. We specifically assess this possibility using the stimulus set from Federmeier et al. (2007), which contains sentences that manipulated the predictability of a final word by designing the sentences to be either strongly or weakly constraining. We therefore sought to test whether it is possible to predict constraint from the sentence embeddings directly to better understand whether and how linguistic uncertainty is encoded in hidden states.

2 Cloze completion dataset

We constructed a cloze completion dataset (Taylor, 1953) to compare one LLM (RoBERTa; (Liu et al., 2019)) to human predictions of final words in Federmeier et al. (2007). This stimulus set contains 282 sentence stimuli that differ in the **constraint** of the sentence, or the degree to which the preamble leads the reader in a specific direction. Broadly, *strongly* constraining sentences consistently lead readers in one semantic direction and the final word (**critical**) is highly predictable; *weakly* constraining sentences are less specific, which we summarize below.

- *Strong Constraint*: Sharon dried the bowls with a **towel**.
- *Weak Constraint*: He always seemed to be interested in looking at the **sky**.

One property of these stimuli is that constraint is partly defined by the predictability of the final word. Cloze probability is defined as the percentage of completions produced by participants that end in a particular final word. For example, if *towel* is guessed by 30% of participants, then its cloze probability is .3. Effects of constraint have sometimes been assessed by categorizing sentences using the cloze probability of the most expected completion, as in Federmeier et al. (2007) [strong constraint: cloze $>$ 67%; weak constraint: cloze $<$ 42%]. Constraint is therefore both a product of the vagueness or specificity of the preamble, and the predictability of the (**bolded**) critical word given the preamble.

The present cloze dataset includes 109,225 word-by-word predictions for all non-initial words from 158 participants recruited from the Prolific platform. All participants self-reported as having acquired American English before age 5 and received

$8 for 30 minutes of their time. Here we are interested in whether the preamble encodes the predictability (constraint) of the final word.

In general, the cloze probability when participants produced the intended final word from Federmeier et al. (2007) was higher when the sentence was strongly constraining (SC; $\hat{\mu}_{SC} = 0.71$) than when it was weakly constraining (WC; $\hat{\mu}_{WC} = 0.21$), $t(266) = 25.1$, $p < .001$. We therefore largely replicated the original divisions of Federmeier et al. (2007). However, the current cloze dataset differs from the original stimulus set in that we are able to leverage the probabilities of all cloze completions to assess uncertainty across these categories. Participants provided more varied responses as evidenced by higher entropy in WC sentences ($\hat{\mu}_{WC} = 0.87$) than in SC ones ($\hat{\mu}_{SC} = 0.36$), $t(200) = 21.8$, $p < .001$, a result we discuss further in Section 3.2. In the next section, we describe our masking procedure for assessing the degree to which cloze probabilities and response entropies correlate with embedding representation-derived measures.

3 Probing the predictability of final words

Given the clear difference in cloze probabilities of critical words in the strongly and weakly constraining stimuli in Section 2, we reasoned that strongly and weakly constraining sentences are relatively easy for participants to distinguish. In this section, however, we sought to test whether the unpredictability of a word as defined by the original cloze labels in Federmeier et al. (2007) is recoverable directly from sentence embeddings, as outlined in Section 2. While this may seem trivial, it is not obvious exactly what factors influence the predictability of a final word – individually or jointly. For example, it is possible that comprehenders rely predominantly on immediately preceding information when completing cloze tasks, but they may also incorporate linguistic properties of words or combinations of words earlier in the sentence (MacDonald and Seidenberg, 2006).

To test whether constraint is recoverable from sentence embeddings, we leveraged the masked language model RoBERTa (Liu et al., 2019), which enabled us to hide the critical words from the model's representation of the sentence and obtain sentence embeddings for a downstream probing model. RoBERTa deviates from human language processing in that it processes the entire sentence simultaneously, rather than incrementally as in recurrent neural networks (Elman, 1990). However, we can present sentences except the final word to RoBERTa, which can mimic any forward predictions and higher-order integration that readers will have done up until that point. Importantly, a masked language model like RoBERTa allows us to mask the final word, and obtain a representation of only the upstream (preamble) part of the sentence.

We then transformed the sentence into a single vector for our classification procedure, taking the original sentence from the Federmeier et al. (2007) stimuli, except we replaced the critical final word with a `<mask>` token. Embedding the sentence using RoBERTa produces a fixed-length vector for each token (roughly, word), from which we computed a sentence embedding vector by averaging all token vectors within each layer, excluding the `<mask>` token. This embedding process produced a $282 \times 13 \times 768$-dimensional matrix. From these embeddings, we then constructed 282 leave-one-out regularized logistic regression probing classifiers (one for each critical sentence) trained on 281 of the sentence embeddings to predict strong (SC) or weak constraint (WC) from the original Federmeier et al. (2007) labels. We then treat the remaining sentence as a test item and obtain a predicted probability of the sentence being strongly constraining.

3.1 Cloze surprisal

In contrast to using raw percentages of completions of the Federmeier et al. (2007) cloze stimuli, we can alternately quantify constraint using either the surprisal of a particular completion (Eq. 1) or estimate entropy (H; Eq. 2) over all K cloze completions:

$$\text{surprisal} = -\log(p(x)) \tag{1}$$

$$H = \sum^{K} p(x) \cdot \log_K(p(x)) \tag{2}$$

If constraint is encoded in both the final resulting sentence and the context, then we expect to see a positive relationship between the model's belief that the sentence is strongly constraining and participants' ability to guess the target word. However, constraint may also be measured using cloze probabilities, or the conditional probability of participants producing a word given a context. In the Federmeier et al. (2007) work, strongly and weakly constraining sentences were designed to have high and low cloze probability completions, respectively. Therefore, we tested for a correlation between linguistic uncertainty as estimated by the cloze probability of the critical word and the predicted probabilities obtained from the classifiers.

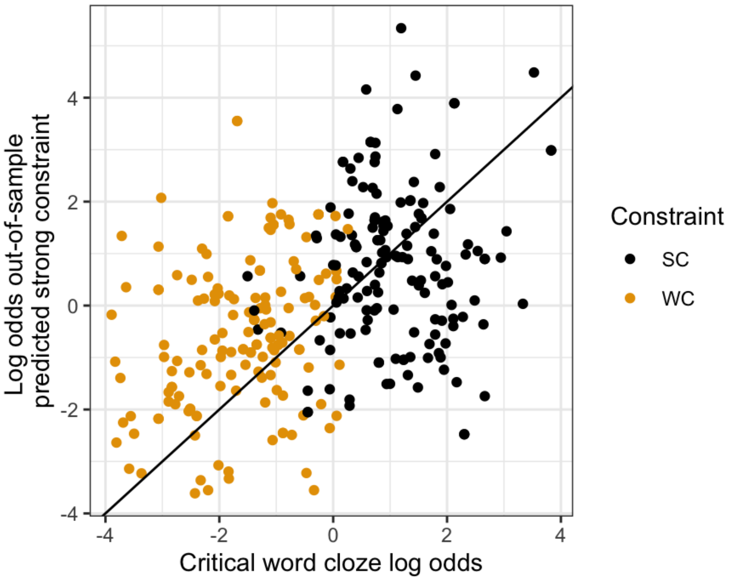

Figure 1: Plotted relationship between critical final word cloze and classifier probability of constraint ($\hat{\rho} = 0.43$, $p < .001$). Line represents perfect correlation.

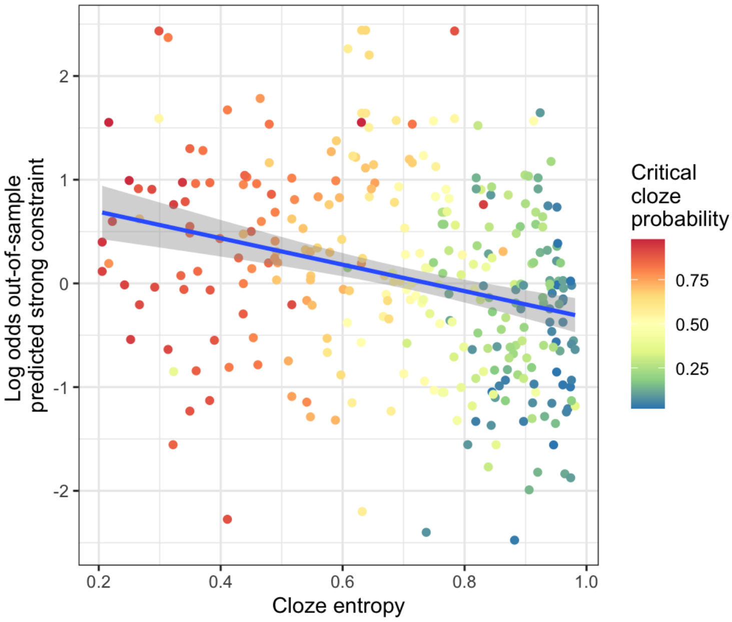

Figure 2: Plotted relationship between final word entropy and classifier probability of constraint. Line represents fitted slope ($\hat{\rho} = -.32$, $p < .001$).

With these predicted probabilities, we then tested for a relationship between the log odds of a predicted SC label as a function of the cloze probability of the final completion and found a strong correlation between the two ($\hat{\rho} = 0.43$, $p < .001$). We plot this relationship in Figure 1.

3.2 Cloze entropy

Like cloze probability, we can also compute the uncertainty of participants' final responses by computing the entropy of the outcomes. This uncertainty captures the intuition that if participants vary in what they expect, then their guesses will be relatively uniformly distributed across many outcomes. Indeed, the weakly constraining sentences in Federmeier et al. (2007) may have been designed to be vague, and thus intended to be completed by many possible valid words. We conducted the same analysis as in the previous section, and found that with greater uncertainty (higher entropy), the model's belief that a sentence was strongly constraining decreased ($\hat{\rho} = -.32$, $p < .001$). We plot this relationship in Figure 2. This strongly suggests that RoBERTa encodes final word uncertainty in a similar way to how it encodes constraint.

4 Discussion

In two experiments we have tested how much context on its own – without knowledge of the final word – can directly encode the predictability of upcoming linguistic information. In contrast to prior work focusing on surprisal, this work leverages experimenter-defined labels (sentential constraint categories) and sentence embeddings derived from the LLM RoBERTa and shows that the model's hidden states directly encode uncertainty about upcoming information. We demonstrated that we are able to train classifiers that can predict the categorical constraint of a sentence and that the model's certainty about the constraint category correlates with the cloze probability of the target word and relatedly the entropy of participants' responses.

These results present an interesting puzzle about how lexical predictability unfolds in human language comprehension. For example, readers build up representations of sentences incrementally as they read through a sentence, though they may read back in a passage or reread some sections of text. In turn, this higher-order representation of the language guides their expectations about upcoming words (Lowder et al., 2018), one aspect of which may be uncertainty or the semantic specificity of predictions that can be made.

In sum, we have presented one of the first attempts at using embeddings instead of computing surprisal values to account for the lexical predictability of words in sentences. We believe that the method outlined here raises several questions about how predictions are launched and how uniformly throughout utterances vagueness or uncertainty is encoded. These questions include topics that are critical from a multiple constraint satisfaction approach (MacDonald and Seidenberg, 2006), such as which words contribute the most toward the predictions of the final words. In future work, we hope to also analyze non-final word uncertainty using similar methods to better understand how cloze probabilities relate to sentence representations as the sentence unfolds. Analyses of attention patterns in LLMs (e.g., Vig and Belinkov, 2019) and masking of specific words may provide some clues to the sources of predictions.

References

Tiwalayo Eisape, Noga Zaslavsky, and Roger Levy. 2020. Cloze distillation: Improving neural language models with human next-word prediction. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 609–619, Online. Association for Computational Linguistics.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Kara D Federmeier, Edward W Wlotko, Esmeralda De Ochoa-Dewald, and Marta Kutas. 2007. Multiple effects of sentential constraint on word processing. *Brain Research*, 1146:75–84.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Matthew W Lowder, Wonil Choi, Fernanda Ferreira, and John M Henderson. 2018. Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive science*, 42:1166–1183.

Maryellen C MacDonald and Mark S Seidenberg. 2006. Constraint satisfaction accounts of lexical and sentence comprehension. In *Handbook of psycholinguistics*, pages 581–611. Elsevier.

Wilson L Taylor. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.