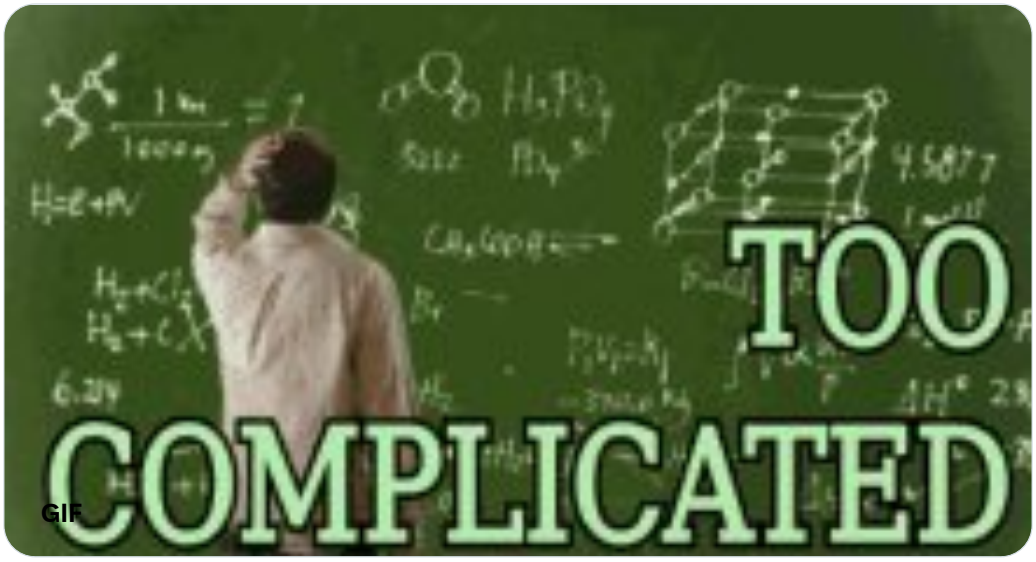
Michael Hahn @mhahn29 · Oct 19, 2022

Take this sentence:

The report that the doctor who the lawyer distrusted annoyed the patient was surprising.

Pretty complicated! It takes some thinking to unpack [the report was surprising, the doctor annoyed the patient, and the lawyer distrusted the doctor].

3/



1 6

Michael Hahn @mhahn29 · Oct 19, 2022

Why and when is syntactic structure tough for humans to comprehend? Prominent perspectives emphasize the role of either

- expectation violation (surprisal) 😲, or

- memory retrieval difficulties 🤔

4/

1 10

Michael Hahn @mhahn29 · Oct 19, 2022

Futrell et al (2020) offer a clever unification of the two ideas: When we process language, we are *always* forgetting the details of what exactly was said.
Difficulty arises when words are surprising given the parts we remember.

5/

1 12

Michael Hahn @mhahn29 · Oct 19, 2022

Futrell et al propose then that we forget material via erasure: randomly deleting previous words, with a bias towards older material.
But they didn't yet specify which material exactly gets erased.
Our current proposal is an extension of this work and addresses this gap.

6/

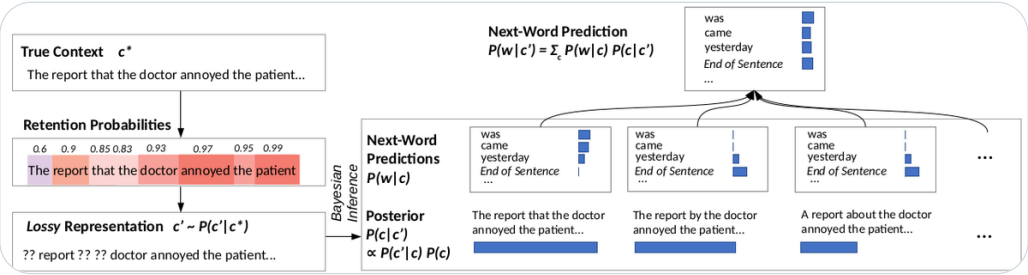
1 6

Michael Hahn @mhahn29 · Oct 19, 2022

We propose a rational tradeoff that maximizes prediction accuracy while minimizing memory cost.

Building on GPT-2, we scale this to arbitrary input while capturing the rich statistical structure of language use.

7/



1 8

Michael Hahn @mhahn29 · Oct 19, 2022

A key prediction follows:
A rational comprehender reconstructs the context based on imperfect memory + what is a-priori likely to have been said.
Hence, expectations will be biased towards contexts with high a-priori probability similar in form to the true contexts.

8/

1 7

Michael Hahn @mhahn29 · Oct 19, 2022

For instance, the context

The report that the doctor annoyed the patient...
[requires a second verb to be a complete sentence]

will compete with variants like

The report *by* the doctor annoyed the patient...
[no verb required]
...
Show more

1 8

Michael Hahn @mhahn29 · Oct 19, 2022

The variant is a complete sentence and no further verb is expected.

If memory is imperfect and such variants have high a-priori probability, comprehension may be derailed when the final verb is encountered, even though it's grammatically necessary.

10/

1 7

Michael Hahn @mhahn29 · Oct 19, 2022

If we manipulate the relative probability of true and variant contexts, we can make the sentence easier or harder to understand.

In English text, "the FACT that" is MUCH more frequent than "the REPORT that", even controlling for baseline frequency of the nouns.

11/

1 8

Michael Hahn @mhahn29 · Oct 19, 2022

Thus, we predict that the following is easier to comprehend (and it is):

The *fact* that the doctor who the lawyer distrusted annoyed the patient was surprising.

because our minds can reconstruct the structurally correct context even when memory is lossy.

12/

1 7

Michael Hahn @mhahn29 · Oct 19, 2022

We crossed a couple of factors that all manipulate the relative probabilities of true and variant contexts

13/

(B) Contexts and Surface-Similar Variants

① Low Embedding Bias, Compatible Verb

	Context c	$P(c)$
True (c^*)	The report that the doctor annoyed the patient...	
Variants	The report **by** the doctor annoyed the patient.	
	The report **about** the doctor annoyed the patient.	
	...	

② Incompatible Verb

	Context c	$P(c)$
True (c^*)	The report that the doctor **cured** the patient...	
Variants	The report **by** the doctor **cured** the patient.	
	The report **about** the doctor **cured** the patient.	
	...	

③ High Embedding Bias

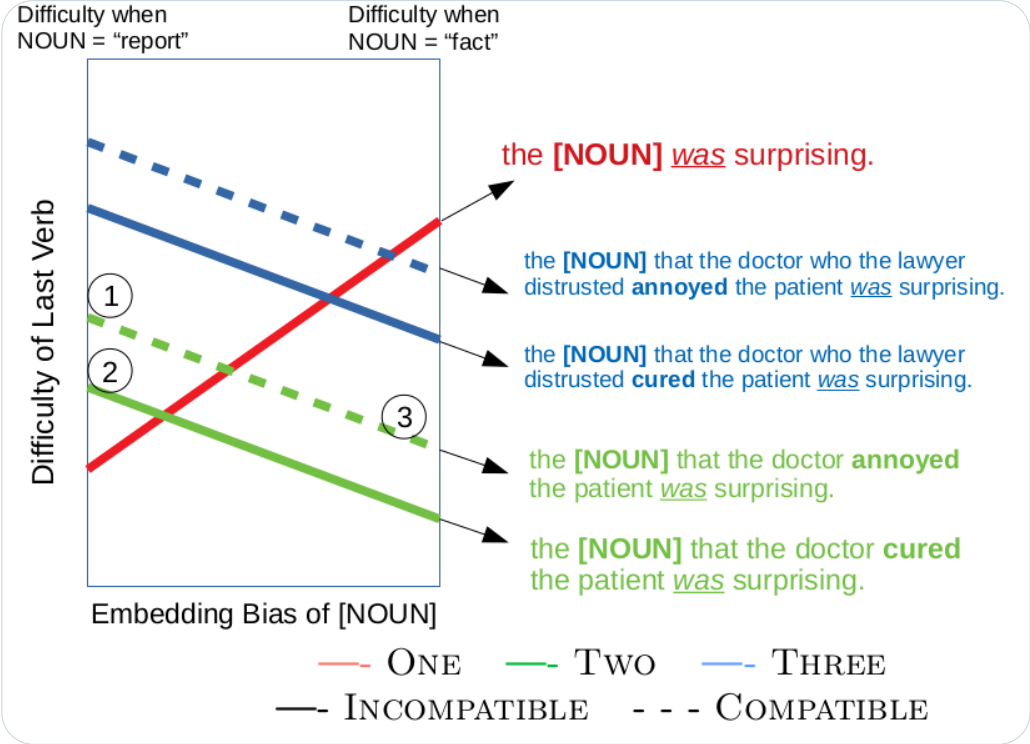
	Context c	$P(c)$
True (c^*)	The **fact** that the doctor annoyed the patient...	
Variants	The **fact of** the doctor annoyed the patient.	
	The **fact about** the doctor annoyed the patient.	

1 6

Michael Hahn @mhahn29 · Oct 19, 2022

and based on this predicted how difficult/disruptive encountering the final verb is

14/

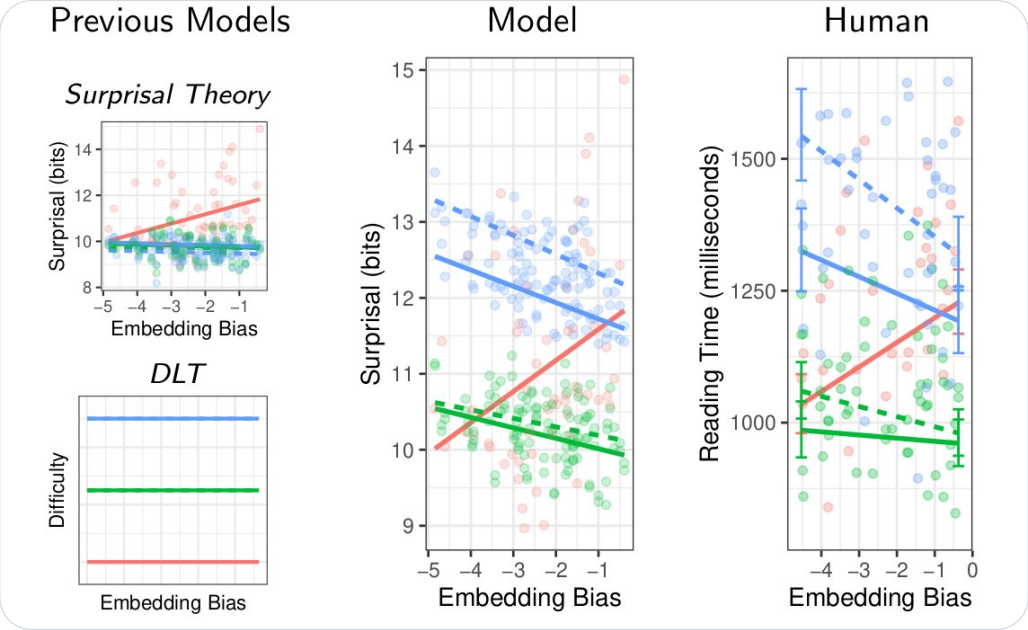


1 5

Michael Hahn @mhahn29 · Oct 19, 2022

Human reading times at the final verb align nicely with the model's prediction.
Both are quite unlike the predictions of existing models.

15/



1 3 9

Michael Hahn @mhahn29 · Oct 19, 2022

We found a corresponding pattern in production:

We asked subjects to complete preambles such as "The report that the doctor who the diplomat...".

This should be followed by three verbs, e.g. “...mistrusted cured the patient was surprising.”.

16/

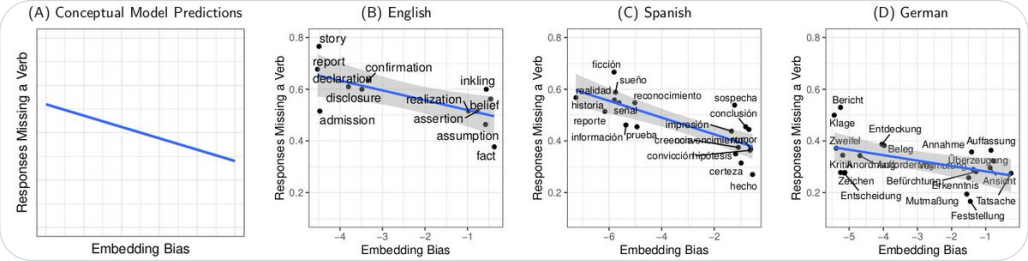
1 6

Michael Hahn @mhahn29 · Oct 19, 2022

This is a hard task: people miss a verb in about 50% of trials.

They are more likely to miss a verb for "fact"-like nouns than for "report"-like nouns, as expected if expectations are derived from noisy memory+language statistics.

17/



2 7

Michael Hahn @mhahn29 · Oct 19, 2022

In this paper, we have probed a couple of ways to manipulate prior probabilities, but there are more. E.g., the theory suggests predictions about how pragmatic reasoning and social expectations could shape syntactic processing.

18/

1 9

Michael Hahn @mhahn29 · Oct 19, 2022

We focused on recursion, but our model makes word-by-word difficulty predictions on arbitrary English input. You can run it on your own stimuli: gitlab.com/m-hahn/resourc...

fin/

gitlab.com
model/compute_surprisal/README.md · main · Mic...
GitLab.com

1 1 14