Attribution & Alignment: Effects of Local Context Repetition on Utterance Production and Comprehension in Dialogue

Aron Molnar[◇] Jaap Jumelet[◁] Mario Giulianelli[◁] Arabella Sinclair[◇]

a.molnar.19@abdn.ac.uk j.w.d.jumelet@uva.nl m.giulianelli@uva.nl arabella.sinclair@abdn.ac.uk

[◇]Department of Computing Science, University of Aberdeen | [◁]Institute for Logic, Language and Computation, University of Amsterdam

Objectives

Understand LM utterance production and comprehension in dialogue settings through psycholinguistics and interpretability techniques.

- Do LMs produce human-like levels of repetition in dialogue?
- What processing mechanisms related to lexical re-use LMs utilize during comprehension?

Repetition is typically penalised when evaluating language model generations. However, it is a key component of dialogue. Humans use ***local*** and ***partner specific*** repetitions; these are preferred by human users and lead to more successful communication in dialogue. We believe that such *joint analysis of model production and comprehension behaviour* can inform the development of *cognitively inspired* dialogue generation systems.

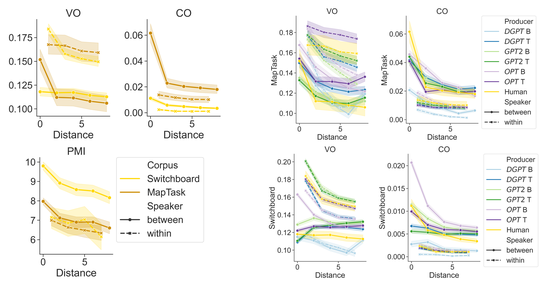
Dialogue Corpora

We define and extract shared constructions—sequences of tokens containing at least two words shared between speakers—from two high quality dialogue corpora.

	Switchboard			MapTask		
	M±Std	Med.	Max	M±Std	Med.	Max
Construction						
Length	2.1 ± 0.4	2.0	5	2.4 ± 0.8	2.0	11
Frequency	3.0 ± 1.2	3.0	6	3.3 ± 1.1	3.0	6
Rep. Dist.	3.6 ± 2.7	3.0	8	3.3 ± 2.7	3.0	8
Incidence	1.6 ± 1.1	1.0	10	2.0 ± 1.1	2.0	8
PMI	6.8 ± 3.4	6.6	11.5	7.2 ± 2.2	7.6	9.6
Utterance						
CO	0.004 ± 0.035	0.0	1.0	0.024 ± 0.13	0.0	2.8
VO	0.13 ± 0.23	0.008	1.0	0.13 ± 0.24	0.0	1.0

Table: Construction properties. Repetition distance (*Rep. Dist.*) measured in utterances.

Repetition



(a) Human properties (b) Human vs. Model properties

Figure: Repetition decay effects for *construction* overlap and *vocabulary* overlap.

We explore locality of repetition effects: the degree to which repetition effects decay with distance between utterances. We differentiate whether a repetition is *between* or *within*-speaker: that is whether a speaker is repeating their interlocutor, or themselves. We also differentiate between repetition at the level of single tokens vs constructions.

Dialogue excerpt - Switchboard. Constructions in ***bold***

A: oh, yeah, yeah, yeah.
B: **in the summer** or like in the easter time, like around now?
A: no, usually **in the summer** time.

Vocabulary Overlap. To compute vocabulary overlap, *VO*, we exclude punctuation, and calculate *VO* as the proportion of words w in the current turn t_c that also appear in a previous turn t_p:

$$VO = \frac{|w_{t_c} \cap w_{t_p}|}{|w_{t_c}|} \tag{1}$$

Construction Repetition. After extracting a shared inventory of constructions for a dialogue, we measure the proportion of repetition of shared constructions C as construction overlap *CO* as:

$$CO = \frac{|C_{t_c} \cap C_{t_p}|}{|w_{t_c}|} \tag{2}$$

Attribution

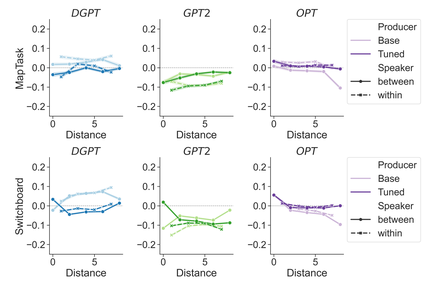


Figure: Relative attribution properties to human utterances over the dialogue context.

We design a measure that aggregates over per-token attributions for a full utterance, returning *relative prediction boosting effects* of tokens within context utterances, speaker label tokens, and the target itself.

We create the feature attribution scores of each token in the input w_i with respect to the prediction of each token in the target utterance w_t:

$$\Phi \in \mathbb{R}^{|w_i| \times |w_t| \times n_{emb}} \tag{3}$$

We sum these scores along the embedding dimension n_{emb}. We sum the Φ matrix along the dimension of the tokens in the target utterance (w_t). We create a single importance score for each individual utterance or turn separator, denoted as a set T_i that contains the indices of the i^{th} utterance:

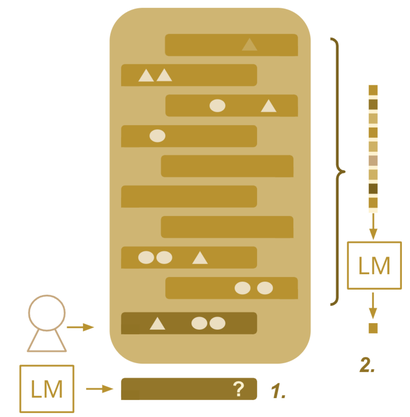
$$\Phi' \in \mathbb{R}^{|T|}, \quad \Phi'_i = \sum_{j \in T_i} \sum_{k} \sum_{l} \Phi_{j,k,l} \tag{4}$$

The scores of Φ' are still unbounded, and can vary greatly between samples and models. We thus normalise the scores by the maximum absolute Φ' score, which maps the scores between -1 and 1, and we then centre the scores around the mean.

$$\Phi'' = \frac{\Phi'}{\max(|\Phi'|)} \tag{5}$$

$$\phi = \Phi'' - \text{mean}(\Phi'') \tag{6}$$

Evaluation procedure



We compare human- vs. LM-produced utterances for a given context. We firstly (1) generate and evaluate the repetition patterns present in a suite of transformer language models given dialogue excerpts of 10 utterances. We then (2) analyse the same models for the salience they assign to the human utterances, to better understand their behaviour when comprehending local repetitions.

Takeaways

- humans repeat word sequences uttered by their dialogue partner locally
- → language models vary in their ability produce similar repetitions
- → while reference-based generation quality metrics correlate with the human-likeness of the repetitions produced, corpus-level metrics fail to capture this important aspect of dialogue quality.
- models assign salience in a local manner when comprehending human utterances
- models assign more salience to utterances containing repetitions in the context