Not wacky vs. definitely wacky: A study of scalar adverbs in pretrained language models.

Isabelle Lorge & Janet Pierrehumbert
Department of Engineering, University of Oxford

Abstract

We carried a systematic study of scalar adverbs, an under-explored class of words with strong logical force. Using three different tasks involving both naturalistic social media data from Reddit and constructed examples, we investigate the extent to which BERT, RoBERTa and GPT-2 exhibit knowledge of these common words. We find that despite capturing some aspects of logical meaning, the models still have obvious shortfalls.

Introduction

- Pretrained language models relying on the distributional assumption may tend to be "blind" to logical meanings
- There are adverbial scales whose items have strict logical relations between them that may not be captured by pretrained language models' training
- Three of these scales are MODALITY, FREQUENCY and DEGREE

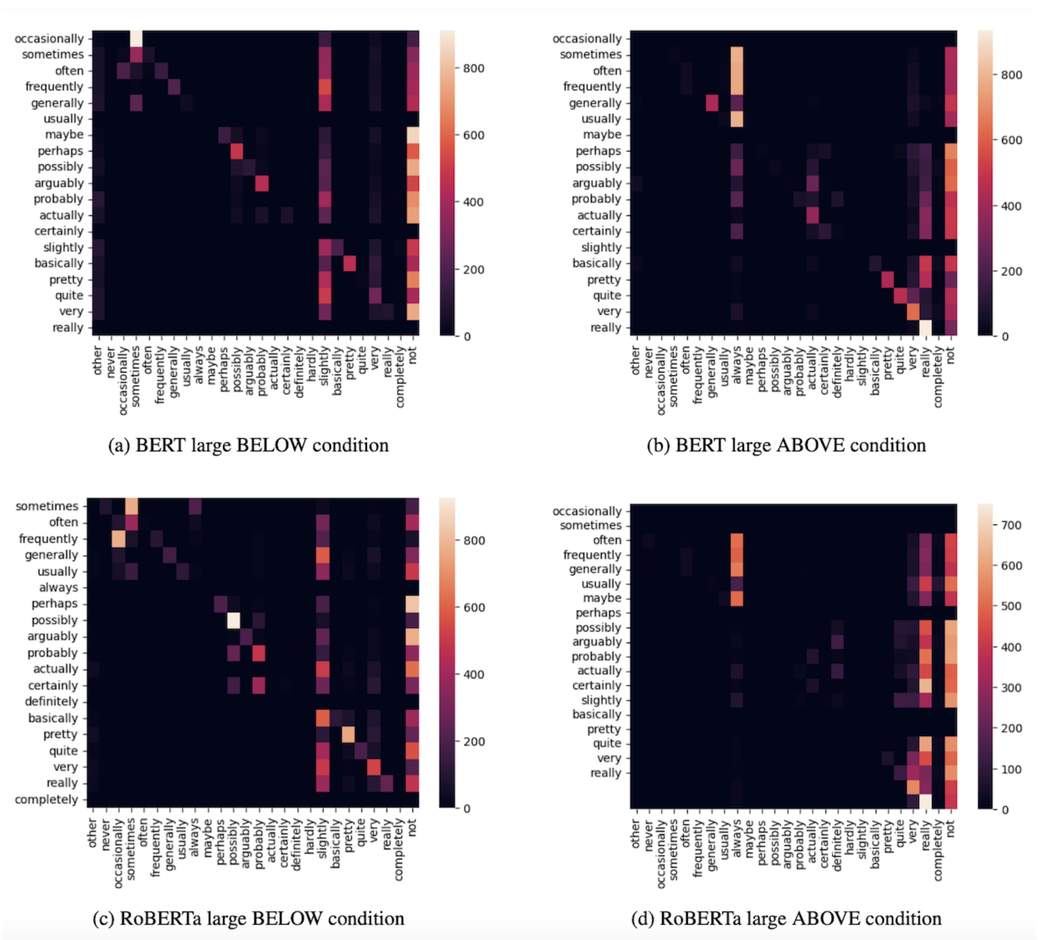
Adverbs

Category	Adverbs
MODALITY (14.8%)	{*maybe, perhaps, possibly*}, *arguably, probably, actually, certainly, definitely*
FREQUENCY (5.3%)	*never, occasionally, sometimes, often, generally, usually, frequently, always*
DEGREE (46.8%)	*hardly, slightly, basically, pretty, quite, very, really, completely*

Method

1. **Ranking task**: Can adverb's embeddings be used to correctly rank them on their respective scales (pairwise accuracy and rank correlations)?
2. **Contextual task**: Are pretrained language models able to use naturalistic context to derive adverbs and do they distinguish them from negation?
3. **Entailment task**: Are pretrained language models able to determine which part of the adverbial scale is licensed in entailment settings?

Entailment task



(a) BERT large BELOW condition

(b) BERT large ABOVE condition

(c) RoBERTa large BELOW condition

(d) RoBERTa large ABOVE condition

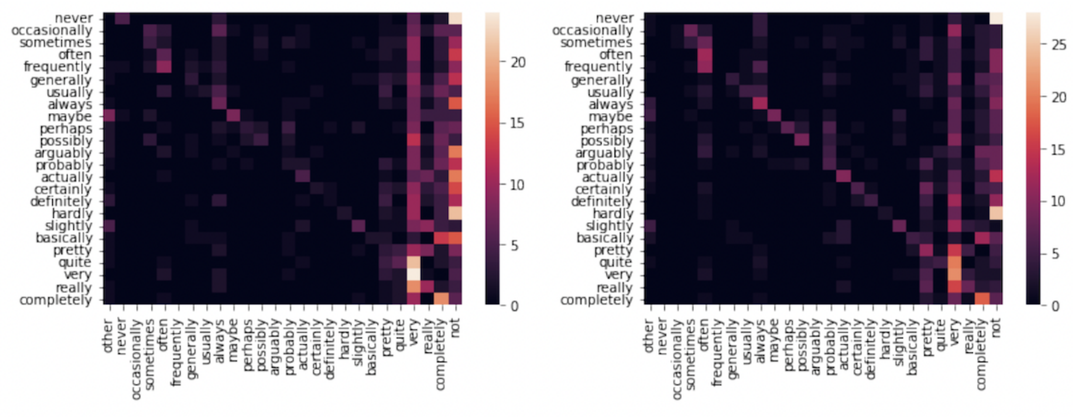
If it is often cold, then it is at least [MASK] cold. (below)	1_below
It is at least [MASK] cold if it is often cold. (below)	2_below
It is often cold so it is at least [MASK] cold. (below)	3_below
It is at least [MASK] cold because it is often cold. (below)	4_below
If it is at most [MASK] cold, then it is not often cold. (below)	5_below
It is not often cold if it is at most [MASK] cold. (below)	6_below
It is at most [MASK] cold so it is not often cold. (below)	7_below
It is not often cold because it is at most [MASK] cold. (below)	8_below

If it is [MASK] blue, then it is at least quite blue. (above)	1_above
It is at least quite blue if it is [MASK] blue. (above)	2_above
It is [MASK] blue so it is at least quite blue. (above)	3_above
It is at least quite blue because it is [MASK] blue. (above)	4_above
If it is at most quite blue, then it is not [MASK] blue. (above)	5_above
It is not [MASK] blue if it is at most quite blue. (above)	6_above
It is at most quite blue so it is not [MASK] blue. (above)	7_above
It not quite blue because it is at most [MASK] blue. (above)	8_above

Ranking task

	Pacc.	Spearman ρ			Kendall τ		
BERT-b	0.60	*f:* 0.68	*m:* 0.77	*d:* 0.32	*f:* 0.52	*m:* 0.66	*d:* 0.23
BERT-l	**0.64**	*f:* **0.78**	*m:* 0.88	*d:* 0.39	*f:* **0.62**	*m:* 0.77	*d:* 0.24
ROBERTA	0.53	*f:* -0.32	*m:* 0.77	*d:* 0.64	*f:* -0.24	*m:* 0.67	*d:* 0.52

Contextual task



(a) BERT large with context

(b) RoBERTa large with context

Target	Context
certainly	You are conflating the issue. Slavery was not moral but it was *[MASK] legal*.
frequently	Doesn't really matter what republicans say, democrats are going to call them racist. Because what Republicans say is *[MASK] racist.*
very	You need verifiable proof. I mean, it's not like saying you're self trained is *[MASK] reputable.*

Conclusions

- **Embeddings contain some limited signal for ranking adverbs within their respective scales but the embeddings space is far from consistent**
- **Models have strong frequency biases, don't distinguish between adverbial scales, and use negation in inappropriate contexts**
- **Models display high variance between logically equivalent contexts and high percentages of negations creating logical contradictions**