Probing Neural Language Models for Human Tacit Assumptions

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme
Department of Computer Science, Johns Hopkins University
{nweir,azpoliak,vandurme}@jhu.edu

Abstract

Humans carry *stereotypic tacit assumptions* (STAs) (Prince, 1978), or propositional beliefs about generic concepts. Such associations are crucial for understanding natural language. We construct a diagnostic set of word prediction prompts to evaluate whether recent neural contextualized language models trained on large text corpora capture STAs. Our prompts are based on human responses in a psychological study of conceptual associations. We find models to be profoundly effective at retrieving concepts given associated properties. Our results demonstrate empirical evidence that stereotypic conceptual representations are captured in neural models derived from semi-supervised linguistic exposure.

Keywords: language models; deep neural networks; concept representations; norms; semantics

Introduction

Recognizing generally accepted properties about concepts is key to understanding natural language (Prince, 1978). For example, if one mentions a bear, one does not have to explicitly describe the animal as having teeth or claws, or as being a predator or a threat. This phenomenon reflects one's held stereotypic tacit assumptions (STAs), i.e. propositions commonly attributed to "classes of entities" (Prince, 1978). STAs, a form of common knowledge (Walker, 1991), are salient to cognitive scientists concerned with how human representations of knowledge and meaning manifest.

As "studies in norming responses are prone to repeated responses across subjects" (Poliak et al., 2018), cognitive scientists demonstrate empirically that humans share assumptions about properties associated with concepts (McRae et al., 2005). We take these conceptual assumptions as one instance of STAs and ask whether recent contextualized language models trained on large text corpora capture them. In other words, do models correctly distinguish concepts associated with a given set of properties? To answer this question, we design fill-in-the-blank diagnostic tests (Figure 1) based on existing data of concepts with corresponding sets of human-elicited properties.

By tracking conceptual recall from prompts of iteratively concatenated conceptual properties, we find that the popular neural language models, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), capture STAs. We observe that RoBERTa consistently outperforms BERT in correctly associating concepts with their defining properties across multiple metrics; this performance discrepancy is consistent with many other language understanding tasks (Wang et al., 2018). We also find that models associate concepts with perceptual categories of properties (e.g. visual) worse than with non-perceptual ones (e.g. encyclopaedic or functional).

Prompt	Model Predictions
A ___ has fur.	dog, cat, fox, ...
A ___ has fur, is big, and has claws.	cat, **bear**, lion, ...
A ___ has fur, is big, has claws, has teeth, is an animal, eats, is brown, and lives in woods.	**bear**, wolf, cat, ...

Figure 1: The concept **bear** as a target emerging as the highest ranked predictions of the neural LM RoBERTa-L (Liu et al., 2019) when prompted with conjunctions of the concept's human-produced properties.

We further examine whether STAs can be extracted *from* the models by designing prompts akin to those shown to humans in psychological studies (McRae et al., 2005; Devereux et al., 2014). We find significant overlap between model and human responses, but with notable differences. We provide qualitative examples in which the models' predictive associations differ from humans', yet are still sensible given the prompt. Such results highlight the difficulty of constructing word prediction prompts that elicit particular forms of reasoning from models optimized purely to predict co-occurrence.

Unlike other work analyzing linguistic meaning captured in sentence representations derived from language models (Conneau et al., 2018; Tenney et al., 2019), we do not fine-tune the models to perform any task; we instead find that the targeted tacit assumptions "fall out" purely from semi-supervised masked language modeling. Our results demonstrate that exposure to large corpora alone, without multi-modal perceptual signals or task-specific training cues, may enable a model to sufficiently capture STAs.

Background

Contextualized Language Models Language models (LMs) assign probabilities to sequences of text. They are trained on large text corpora to predict the probability of a new word based on its surrounding context. Uni-

directional models approximate for any text sequence $w = [w_1, w_2, \ldots w_N]$ the factorized left-context probability $p(w) = \prod_{i=1}^{N} p(w_i \mid w_1 \ldots w_{i-1})$. Recent neural *bi-directional* language models estimate the probability of an intermediate 'masked out' token given both left and right context; this task is colloquially "masked language modelling" (MLM). Training in this way produces a probability model that, given input sequence w and an arbitrary vocabulary word v, predicts the distribution $p(w_i = v \mid w_1, \ldots w_{i-1}, w_{i+1}, \ldots w_n)$. When neural bi-directional LMs trained for MLM are subsequently used as contextual encoders,[1] performance across a wide range of language understanding tasks greatly improves.

We investigate two recent neural LMs: **B**i-directional **E**ncoder **R**epresentations from **T**ransformers (BERT) (Devlin et al., 2019) and **R**obustly **o**ptimized **BERT** **a**pproach (ROBERTA) (Liu et al., 2019). In addition to the MLM objective, BERT is trained with an auxiliary objective of next-sentence prediction. BERT is trained on a book corpus and English Wikipedia. Using an identical neural architecture, ROBERTA is trained for purely MLM (no next-sentence prediction) on a much larger dataset with words masked out of larger input sequences. Performance increases ubiquitously on standard NLU tasks when BERT is replaced with ROBERTA as an off-the-shelf contexual encoder.

Probing Language Models via Word Prediction Recent research employs word prediction tests to explore whether contextualized language models capture a range of linguistic phenomena, e.g. syntax (Goldberg, 2019), pragmatics, semantic roles, and negation (Ettinger, 2020). These diagnostics have psycholinguistic origins; they draw an analogy between the "fill-in-the-blank" word predictions of a pre-trained language model and distribution of aggregated human responses in cloze tests designed to target specific sentence processing phenomena. Similar tests have been used to evaluate how well these models capture symbolic reasoning (Talmor et al., 2019) and relational facts (Petroni et al., 2019).

Stereotypic Tacit Assumptions Recognizing associations between concepts and their defining properties is key to natural language understanding and plays "a critical role in language both for the conventional meaning of utterances, and in conversational inference" (Walker, 1991). *Tacit assumption* (TAs) are commonly accepted beliefs about specific entities (*Alice has a dog*) and *stereotypic* TAs (STAs) pertain to a generic concept, or a class of entity (*people have dogs*) (Prince, 1978). While held by individuals, STAs are generally agreed upon and are vital for reflexive reasoning and pragmatics; Alice might tell Bob 'I have to walk my dog!,' but she does not need to say "I am a person, and people have dogs, and dogs need to be walked, so I have to walk my dog!" Comprehending STAs allows for generalized recognition of new categorical instances, and facilitates learning *new* categories (Lupyan et al., 2007), as shown in early word learning by children (Hills et al., 2009). STAs are not explicitly facts.[2] Rather, they are sufficiently probable assumptions to be associated with concepts by a majority of people. A partial inspiration for this work was the observation by Van Durme (2010) that the concept attributes most supported by peoples' *search engine query logs* (Pasca & Van Durme, 2007) were strikingly similar to examples of STAs listed by Prince. That is, there is strong evidence that the beliefs people hold about particular conceptual attributes (e.g. "countries have kings"), are reflected in the aggregation of their most frequent search terms ("what is the name of the king of France?").

Our goal is to determine whether contextualized language models exposed to large corpora encode associations between concepts and their tacitly assumed properties. We develop probes that specifically test a model's ability to recognize STAs. Previous works (Rubinstein et al., 2015; Sommerauer & Fokkens, 2018; Da & Kasai, 2019) have tested for similar types of stereotypic beliefs; they use supervised training of probing classifiers (Conneau et al., 2018) to identify concept/attribute pairs. In contrast, our word prediction diagnostics find that these associations *fall out* of semi-supervised LM pretraining. In other words, the neural LM inducts STAs as a byproduct of learning co-occurrence without receiving explicit cues to do so.

Probing for Stereotypic Tacit Assumptions

Despite introducing the notion of STAs, Prince (1978) provides only a few examples. We therefore draw from other literature to create diagnostics that evaluate how well a contexualized language model captures the phenomenon. Semantic feature production norms, i.e. properties elicited from human subjects regarding generic concepts, fall under the category of STAs. Interested in determining "what people know about different things in the world,"[3] McRae et al. (2005) had human subjects list properties that they associated with individual concepts. When many people individually attribute the same properties to a specific concept, collectively they provide STAs. We target the elicited properties that were most often repeated across the subjects.

Prompt Design We construct prompts for evaluating STAs in LMs by leveraging the CSLB Concept Property Norms (Devereux et al., 2014), a large extension of the McRae study that contains 638 concepts each

[1] That is, when used to obtain contextualized representations of words and sequences.

[2] E.g., "countries have presidents" does not apply to *all* countries.

[3] Wording taken from instruction shown to participants—as shown in Appendix B of McRae et al. (2005)

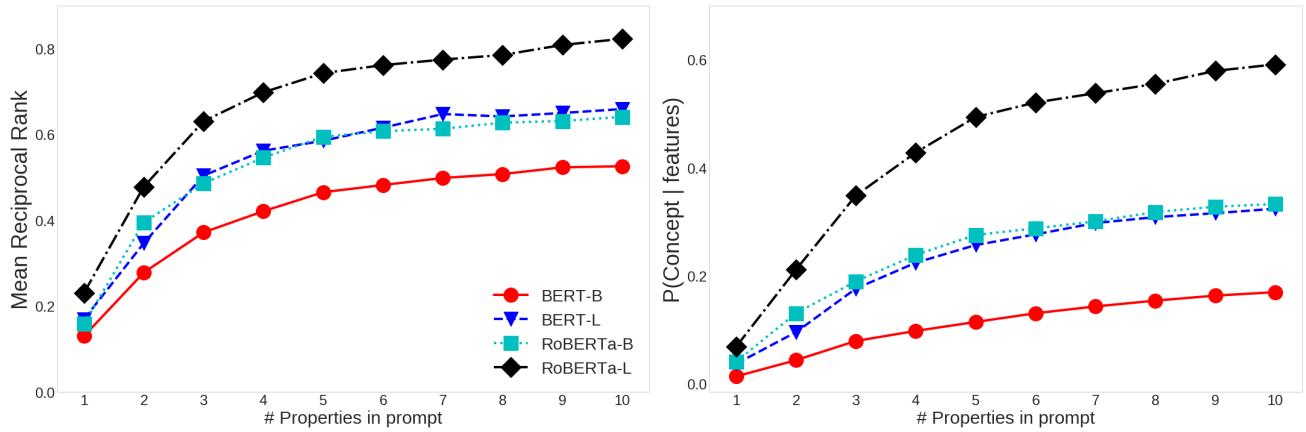

Figure 2: Results from neural LM concept retrieval diagnostic. Mean reciprocal rank and assigned probability of correct concept word sharply increase with the number of conjunctive properties in the prompt.

linked with roughly 34 associated properties. The fill-in-the-blank prompts are natural language statements in which the target concept associated with a set of human-provided properties is the missing word . If LMs accurately predict the missing concept, we posit that they encode the given STA set. We iteratively grow prompts by appending conceptual properties into a single compound verb phrase (Figure 1) until the verb phrase contains 10 properties. Since we test for 266 concepts, this process creates a total of 2,660 prompts.[4] Devereux et al. (2014) record production frequencies (PF) enumerating how many people produced each property for a given concept. For each concept, we select and append the properties with the highest PF in decreasing order. Iteratively growing prompts enables a *gradient of performance* - we observe concept retrieval given few "clue" properties and track improvements as more are given.

Probing Method Prompts are fed as toknized sequences to the neural LM encoder with the concept token replaced with a `[MASK]`. A softmax is taken over the final hidden vector extracted from the model at the index of the masked token to obtain a probability distribution over the vocabulary of possible words. Following Petroni et al. (2019), we use a pre-defined, case-sensitive vocabulary of roughly 21K tokens to control for the possibility that a model's vocabulary size influences its rank-based performance.[5] We use this probability distribution to obtain a ranked list of words that the model believes should be the missing *t* token. We evaluate the BASE (-B) and LARGE (-L) cased models of BERT and ROBERTA.

Evaluation Metrics We use mean reciprocal rank (MRR), or $1/\text{rank}_{\text{LM}}(\text{target concept})$, a metric more sensitive to fine-grained differences in rank than other common retrieval metrics such as recall. We track the predicted rank of a target concept from relatively low ranks given few 'clue' properties to much higher ranks as more properties are appended. MRR above 0.5 for a test set indicates that a model's top 1 prediction is correct in a majority of examples. We also report the overall probability the LM assigns to the target concept regardless of rank. This allows us to measure model *confidence* beyond empirical task performance.

Results

Figure 2 displays the results. When given just one property, ROBERTA-L achieves a MRR of 0.23, indicating that the target concept appears on average in the model's top-5 fill-in predictions (over the whole vocabulary). The increase in MRR and model confidence (y-axis) as properties are iteratively appended to prompts (increasing x-axis) demonstrates that the LMs more accurately retrieve the missing concept when given more associated properties. MRR steeply increases for all models as properties are added to a prompt, but we find less stark improvements after the first four or five. The LARGE models consistently outperform their BASE variants under both metrics, as do ROBERTAs over the BERTs of the same size. ROBERTA-B and BERT-L perform interchangeably. Notably, ROBERTA-L achieves a higher performance on both metrics when given just 4 'clue' properties than any other model when provided with all 10. ROBERTA-L assigns *double* the target probability at 10 properties than that of the next best model (ROBERTA-B). Thus, ROBERTA-L is profoundly more confident in its *correct* answers than any

[4]Because LMs are highly sensitive to the 'a/an' determiner preceding a masked word e.g. LMs far prefer to complete *"**A** ___ buzzes,"* with "bee," but prefer e.g. "insect" to complete *"**An** ___ buzzes."*, a task issue noted by Ettinger (2020). We remove examples containing concepts that begin with vowel sounds. A prompt construction that simultaneously accepts words that start with both vowels and consonants is left for future work.

[5]The vocabulary is constructed from the unified intersection of those used to train BERT and ROBERTA. We omit concepts that are not contained within this intersection.

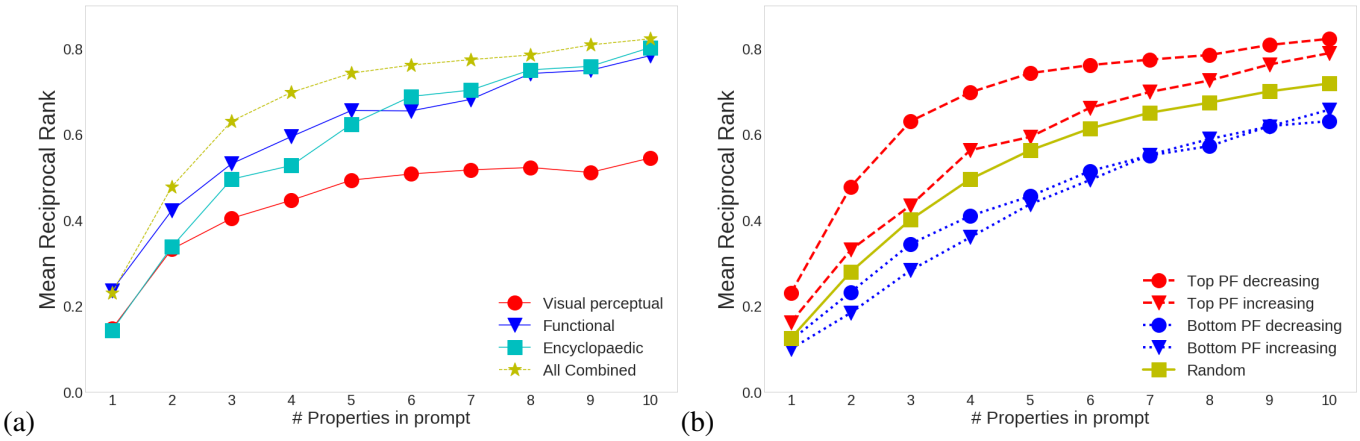

Figure 3: (a) Comparison of RoBERTa-L's performance given only properties from each category versus all combined. (b) RoBERTa-L performance given the property sets with the top vs bottom production frequencies (PF) ordered in increasing vs decreasing PF. Plotted against a randomly sampled and reordered baseline.

other model. However, that all models achieve at least between .5 and .85 MRR conditioned on 10 properties illustrates the effectiveness of all considered models in identifying concepts given STA sets.

Qualitative Analysis We find that model predictions are nearly always grammatical and semantically sensible. Highly-ranked incorrect answers generally apply to a subset of the conjunction of properties, or are correct at an intermediate iteration but become precluded by subsequently appended properties. [6] We note that an optimal performance may not be perfect; not all prompts uniquely identify the target concept, even at 10 properties. [7] However, models still perform nearly as well as could be expected given the ambiguity.

Properties Grouped by Category To measure whether the the *type* of property affects the ability of LMs to retrieve a concept, we create additional prompts that only contain properties of specific categories as grouped by Devereux et al. (2014): visual perceptual (bears have fur), functional (eat fish), and encyclopaedic (are found in forests).[8]

Figure 3a shows that RoBERTa-L performs interchangeably well given just encyclopedic or functional type properties. BERT (not shown) shows a similar overall pattern, but it performs slightly better given encyclopedic properties than functional. Perceptual properties are overall less helpful for models to distinguish concepts than non-perceptual. This may be the product of category specificity; while perceptual properties are produced by humans nearly as frequently as non-perceptual, the average perceptual property is assigned to nearly twice as many CSLB concepts as the average non-perceptual (6 to 3). However, the empirical finding coheres with previous conclusions that models that learn from language alone lack knowledge of perceptual features (Collell & Moens, 2016; Lucy & Gauthier, 2017).

Selecting and Ordering Prompts When designing the probes, we selected and appended the 10 properties with the highest production frequencies (PF) in decreasing order. To investigate whether these selection and ordering choices affect a model's performance in the retrieval task, we compare the top-PF property selection method with an alternative selection criterion using the *bottom*-PF properties. For both selection methods, we compare the decreasing-PF ordering with a reversed, increasing-PF order. We compare the resulting 4 evaluations against a random baseline that measures performance using a random permutation of a randomly-selected set of properties.[9]

Figure 3b compare the differences in performance. Regardless of ordering, the selection of the top (bottom)-PF features improves (reduces) model performance relative to the random baseline. Ordering by decreasing PF improves performance over the opposite direction by up to 0.2 for earlier sizes of property conjunction, but the two strategies converge in performance for larger sizes. This indicates that the selection and ordering criteria of the properties may matter when adding them to prompts. The properties with lower PF are correspondingly less beneficial for model performance. This suggests that assumptions that are less stereotypic—that is, highly salient to fewer humans—are less well captured by the LMs.

[6] E.g. *tiger* and *lion* are correct for *'A ___ has fur, is big, and has claws'* but reveal to be incorrect with the appended *'lives in woods'*

[7] E.g. the properties of *buffalo* do not distinguish it from *cow*.

[8] We omit the categories "other perceptual" (bears growl) and "taxonomic" (bears are animals), as few concepts have more than 2-3 such properties.

[9] The random baseline's performance is averaged over 5 random permutations of 5 random sets for each concept.

Eliciting Properties from Language Models

We have found that neural language models capture to a high degree the relationship between human-produced sets of stereotypic tacit assumptions and their associated concepts. Can we use the LMs to *retrieve* the conceptual properties under the same type of setup used for human elicitation? We design prompts to replicate the "linguistic filter" (McRae et al., 2005) through which the human subjects conveyed conceptual assumptions.

In the human elicited studies, subjects were asked to list properties that would complete "{concept} {relation}..." prompts in which the relation could take on[10] one of four fixed phrases: *is*, *has*, *made of*, and *does*. We mimic this protocol using the first three relations[11] and compare the properties predicted by the LMs to the corresponding human response sets. Examples of this protocol are shown in Table 1.

Comparing LM Probabilities with Humans We can consider the listed properties as samples from a fuzzy notion of a human STA *distribution* conditioned on the concept and relation. These STAs reflect how humans codify their probabilistic beliefs about the world. What a subject writes down about the 'dog' concept reflects what that subject believes from their experience to be sufficiently ubiquitous, i.e. extremely probable, for all 'dog' instances. The dataset also portrays a distribution *over* listed STAs. Not all norms are produced by all participants given the same concepts and relation prompts; this reflects how individuals hold different sets of STAs about the same concept. Through either of these lenses, we can speculate that the human subject produces the sample e.g. 'fur' from some $p(\text{STA} \mid \text{concept} = \textit{bear},\ \text{relation} = \textit{has})$. [12] We can consider our protocol to be sampling from a LM approximation of such a conditional distribution.

Limits to Elicitation Asking language models to list properties via word prediction is inherently limiting, as the models are not primed to specifically produce *properties* beyond whatever cues we can embed in the context of a sentence. In contrast, human subjects were asked directly "What are the properties of X?" (Devereux et al., 2014). This is a highly semantically constraining question that cannot be directly asked of an off-the-shelf language model.

The phrasing of the question to humans also has implications regarding salience: when describing a dog, humans would rarely, if never, describe a dog as being *"larger than a pencil"*, even though humans are "capable of verifying" this property (McRae et al., 2005). Even if they do produce a property as opposed to an alternative lexical completion, it may be unfair to expect language models to replicate how human subjects prefer to list properties that distinguish and are salient to a concept (e.g. *'goes moo'*) as opposed to listing properties that apply to many concepts (e.g. *'has a heart'*). Thus, comparing properties elicited by language models to those elicited by humans is a challenging endeavour. Anticipating this issue, we prepend the phrase 'Everyone knows that' to our prompts. They therefore take the form shown in the left column of Table 1. For the sake of comparability, we evaluate the models' responses against only the human responses that fit the same syntax. We also remove human-produced properties with multiple words following the relation (e.g. '*is found in forests*') since the contextualized LMs under consideration can only predict a single missing word. Our method produces a set of between 495 and 583 prompts for each of the relations considered.

Context	Human		RoBERTa-L	
	Response	**PF**	**Response**	p_{LM}
(Everyone knows that) a bear has ___ .	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02
(Everyone knows that) a ladder is made of ___ .	metal	25	wood	.33
	wood	20	steel	.08
	plastic	4	metal	.07
	aluminum	2	aluminum	.03
	rope	2	concrete	.03

Table 1: Example concept/relation prompts with resulting human and RoBERTa-L responses (and corresponding production frequencies and LM probabilities, resp.). Portions of context prompts encased in () were only shown to the model, not human.

Results We use the information retrieval metric mean average precision (mAP) for ranked sequences of predictions in which there are multiple correct answers. We define mAP here given n test examples:

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{|\text{vocab}|} P_i(j)\Delta r_i(j)$$

where $P_i(j)$ = precision@j and $\Delta r_i(j)$ is the change in recall from item $j-1$ to j for example i. We report mAP on prediction ranks over a LM's entire vocabulary ($\text{mAP}_{\text{VOCAB}}$), but also over a much smaller vocabulary (mAP_{SENS}) comprising the set of human completions that fit the given prompt syntax *for all concepts in the study*. This follows the intuition that responses given for a set of concepts are likely *not* attributes of the other

[10] Selected at the discretion of the subject via a drop-down menu.

[11] We do not investigate the *does* relation or the open-ended *"..."* relation, because the resulting human responses are not easily comparable with LM predictions using template-based prompts. We construct prompts using *is a* and *has a* for broader dataset coverage.

[12] This formulation should be taken with a grain of salt; the subject is given all relation phrases at once and has the opportunity to fill out as many (or few) completions as she deems salient, provided that in combination there are at least 5 total properties listed.

Relation	|Data|	Metric	Bb	Bl	Rb	Rl
is	583	$\mathbf{mAP}_{\text{VOCAB}}$	.081	.080	.078	**.190**
		$\mathbf{mAP}_{\text{SENS}}$	.131	.132	.105	**.212**
		$\rho_{\textbf{Human PF}}$	.062	.100	.062	**.113**
is a	506	$\mathbf{mAP}_{\text{VOCAB}}$	.253	.318	.266	**.462**
		$\mathbf{mAP}_{\text{SENS}}$	.393	.423	.387	**.559**
		$\rho_{\textbf{Human PF}}$	.226	**.389**	.385	.386
has	564	$\mathbf{mAP}_{\text{VOCAB}}$	.098	.043	.151	**.317**
		$\mathbf{mAP}_{\text{SENS}}$	.171	.138	.195	**.367**
		$\rho_{\textbf{Human PF}}$	.217	.234	.190	**.316**
has a	537	$\mathbf{mAP}_{\text{VOCAB}}$	.202	.260	.136	**.263**
		$\mathbf{mAP}_{\text{SENS}}$	.272	.307	.208	**.329**
		$\rho_{\textbf{Human PF}}$	.129	.153	.174	**.209**
made of	495	$\mathbf{mAP}_{\text{VOCAB}}$	.307	.328	.335	**.503**
		$\mathbf{mAP}_{\text{SENS}}$	.324	.339	.347	**.533**
		$\rho_{\textbf{Human PF}}$	.193	.182	.075	**.339**

Table 2: Mean average precision and Spearman ρ with human PF for LM prediction of properties given concept/relation pairs. **B** and **R** indicate BERT and ROBERTA, **b** and **l** indicate -BASE and -LARGE.

concepts, and models should be sensitive to this discrepancy. While mAP measures the ability to distinguish the *set*[13] of correct responses from incorrect responses, we also evaluate probability assigned *among* the correct answers by computing average Spearman's ρ between human production frequency and LM probability.

Results using these metrics are displayed in Table 2. We find that ROBERTA-L outperforms the other models by up to double mAP. No model's rank ordering of correct answers correlates particularly strongly with human production frequencies. When we narrow the models' vocabulary to include only the property words produced by humans for a given syntax, we find that performance (mAP_{SENS}) increases ubiquitously.

Qualitative Analysis Models generally provide coherent and grammatically acceptable completions. Most outputs fall under the category of 'verifiable by humans,' which as noted by McRae et al. could be listed by humans given sufficient instruction. We observe properties that apply to the concept but are not in the dataset [14] and properties that apply to senses of a concept that were not considered in the human responses. [15] We find that some prompts are not sufficently syntactically constraining, and license non-nominative completions. The relation *has* permits past participle completions (e.g. 'has arrived') along with the targeted nominative attributes ('has wheels'). We also find that models idiosyncratically favor specific words regardless of the concept, which can lead to unacceptable completions.[16]

[13] Invariant to order of correct answers.

[14] E.g. '*hamsters are real*' and '*motorcycles have horsepower*'.

[15] While human subjects list only properties of the object *anchor*, LMs also provide properties of a television anchor.

[16] ROBERTA-B often blindly produces '*has legs*', the two BERT models predict that nearly all concepts are '*made of wood*,' and all models except ROBERTA-L often produce '*is dangerous*.'

Prince Example	ROBERTA-L
A **person** has parents, siblings, relatives, a home, a pet, a car, a spouse, a job. ,	person [.73], child [.1], human [.04], family [.03], kid [.02]
A country has a **leader**, a **duke**, **borders**, a **president**, a **queen**, **citizens**, **land**, a **language**, and a **history**.	constitution [.23], history [.07], culture [.07], soul [.04], budget [.03], border [.03], leader [.03], currency [.02], population [.02]

Table 3: ROBERTA-L captures Prince's own exemplary STAs (target completions bolded), as shown by predictions of both concept and properties (associated probability in brackets).

Effect of Prompt Construction We investigate the extent to which our choice of lexical framing impacts model performance by ablating the step in which "everyone knows that" is prepended to the prompt. We find a relatively wide discrepancy in effects; with the lessened left context, models perform on average .05 and .1 mAP worse on the *is* and *has* relations respectively, but perform .06 and .01 mAP better on *is a* and *has a*. Notably, ROBERTA-L sees a steep drop in performance on the *has* relation, losing nearly .3 mAP. We observe that models exhibit highly varying levels of instability given the choice of context. This highlights the difficulty in constructing prompts that effectively target the same type of lexical response from any arbitrary bidirectional LM.

Capturing Prince's STAs

We return to Prince (1978) to investigate whether neural language models, which we have found to capture STAs elicited from humans by McRae, do so as well for what she envisioned. Prince lists some of her *own* STAs about the concepts *country* and *person*. We apply the methodologies of the previous experiments and show the resulting conceptual recall and feature productions in Table 3. We find significant overlap in both directions of prediction. Thus, the exact examples of basic information about the world that Prince considers core to discourse and language processing are clearly captured by the neural LMs under investigation.

Conclusion

We have explored whether the notion owing to Prince (1978) of the stereotypic tacit assumption (STA), a type of background knowledge core to natural language understanding, is captured by contexualized language modeling. We developed diagnostic experiments derived from human subject responses to a psychological study of conceptual representations and observed that recent contextualized LMs trained on large corpora may indeed capture such important information. Through word prediction tasks akin to human cloze tests, our results provide a lens of quantitative and qual-

itative exploration of whether BERT and RoBERTa capture concepts and associated properties. We illustrate that the conceptual knowledge elicited from humans by Devereux et al. (2014) is indeed contained within an encoder; when a speaker may mention something that *'flies'* and *'has rotating blades,'* the LM can infer the description is of a *helicopter*. We hope that our work serves to further research in exploring the extent of semantic and linguistic knowledge captured by contextualized language models.

Acknowledgements

This work was supported in part by DARPA KAIROS (FA8750-19-2-0034). The views and conclusions contained in this work are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

Collell, G., & Moens, M.-F. (2016). Is an image worth more than a thousand words? on the fine-grain semantic differences between visual and linguistic representations. In *COLING*.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *ACL*.

Da, J., & Kasai, J. (2019). Cracking the Contextual Commonsense Code: Understanding Commonsense Reasoning Aptitude of Deep Contextual Representations. In *First Workshop on Commonsense Inference in Natural Language Processing*.

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The centre for speech, language and the brain (CSLB) concept property norms. *Behavior Research Methods*, *46*(4).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *TACL*, *8*, 34–48.

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Categorical Structure among Shared Features in Networks of Early-learned Nouns. *Cognition*, *112*(3), 381–396.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lucy, L., & Gauthier, J. (2017). Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *First Workshop on Language Grounding for Robotics*.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077-1083.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559.

Pasca, M., & Van Durme, B. (2007). What you seek is what you get: Extraction of class attributes from query logs. In *IJCAI*.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language models as knowledge bases? In *EMNLP*.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018). Hypothesis only baselines in natural language inference. In *Starsem*.

Prince, E. F. (1978). On the function of existential presupposition in discourse. In *Chicago Linguistic Society* (Vol. 14, pp. 362–376).

Rubinstein, D., Levi, E., Schwartz, R., & Rappoport, A. (2015). How well do distributional models capture different types of semantic knowledge? In *ACL*.

Sommerauer, P., & Fokkens, A. (2018). Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. In *BlackboxNLP*.

Talmor, A., Elazar, Y., Goldberg, Y., & Berant, J. (2019). oLMpics – On what Language Model Pre-training Captures. *arXiv preprint arXiv:1912.13283*.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., ... Pavlick, E. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. In *ICLR*.

Van Durme, B. (2010). *Extracting Implicit Knowledge from Text*. Unpublished doctoral dissertation, University of Rochester.

Walker, M. A. (1991). *Common Knowledge: A Survey*. University of Pennsylvania.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *EMNLP*.