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 propositional assumptions about generic concepts that are crucial for understanding the semantics of natural language. We ask whether recent powerful language encoders trained on large text corpora capture such *stereotypic tacit assumptions* (STAs) (Prince, 1978). We construct a set of word prediction prompts to evaluate whether recent contextualized language models, BERT and RoBERTa, capture STAs elicited from humans in a psychological study of human conceptual associations. We find the latter model to be profoundly effective at both retrieving concepts given their associated properties and producing properties associated with concepts. Our results demonstrate empirical evidence that stereotypic conceptual representations are captured in neural models derived from linguistic exposure.

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

Introduction

Recognizing generally accepted properties about concepts are 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 being a predator or a threat. This phenomenon reflects 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, Naradowsky, Haldar, Rudinger, & Van Durme, 2018), cognitive scientists demonstrate that humans share assumptions about properties associated with concepts (McRae, Cree, Seidenberg, & McNorgan, 2005). We ask whether contextualized language models trained on large corpora capture STAs. In other words, do these models correctly distinguish concepts associated with a given set of properties? To answer this question, we design cloze tests (Figure 1) based on existing data of human-elicited concepts with corresponding sets of properties.

We find that popular language models trained on large corpora, e.g. BERT (Devlin, Chang, Lee, & Toutanova, 2018) and RoBERTa (Liu et al., 2019), capture STAs. Furthermore, RoBERTa consistently outperforms BERT in correctly associating concepts with their defining properties across multiple metrics. Our analyses indicate that these models associate concepts with different categories of properties better than with other categories of properties. Furthermore, we provide qualitative examples where the models' associations differ from the human-elicited associations, yet are still sensible. Unlike other work analyzing linguistic meaning captured in language models, we do not fine-tine the language models to the type of reasoning we evaluate for. Therefore, our results demonstrate that exposure to large corpora alone, without multi-modal perceptual signals, may enable a model to sufficiently capture STAs.

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 of RoBERTa's ranked predictions given a conjunction of human-produced properties.

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 the preceding sequence. learn a distribution. Unidirectional models approximate for any sequence $w = [w_1, w_2, \ldots w_N]$ the factorized distribution $p(w) = \prod_{i=1}^{N} p(w_i \mid w_1 \ldots w_{i-1})$. Recent neural *bi-directional* language models do not have a well-formed probability of entire sequences as they are trained to estimate the probability of an intermediate token that has been removed from a sequence. Given input sequence w with a randomly-selected word $w_i, 1 \leq i \leq N$, the contextual LM is typically trained to predict the distribution $\Pr(w_i \mid w_1, \ldots w_{i-1}, w_{i+1}, \ldots w_n)$. When these neural bi-directional models pre-trained on larger corpora are used as contextual encoders, performance across a wide range of natural language understanding tasks drastically improves.

We investigate two recent neural language models: **Bi**directional **E**ncoder **R**epresentations from **T**ransformers (BERT) (Devlin et al., 2018) and **R**obustly **o**ptimized **BERT** **a**pproach (RoBERTa) (Liu et al., 2019). BERT In addition to the objective of filling in randomly-masked words, BERT

is trained with an auxiliary objective of next-sentence prediction. BERT was trained on the BookCorpus (Zhu et al., 2015) and English Wikipedia. Using the identical neural architecture, ROBERTA was not trained with the next-sentence auxiliary objective but was trained on more data and larger input sequences. Performance increased on standard NLU datasets when BERT was replaced with ROBERTA as an off-the-shelf contexualized encoder.

Probing Contextualized Language Models 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, 2019). These diagnostics have psycholinguistic origins; they draw an analogy between the "fill-in-the-blank" word predictions of a pre-trained language model and distributions of aggregated human responses in cloze tests. Similar tests have been used to evaluate how well these models capture symbolic reasoning (Talmor, Elazar, Goldberg, & Berant, 2019) and relational facts (Petroni et al., 2019). We also probe these models with cloze tests.

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 objects (*Alice has a dog*) and *stereotypic* TAs (STAs) pertain to a generic concept, or a class of objects (*people have dogs*) (Prince, 1978). 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, Rakison, & McClelland, 2007), as shown in early word learning of young children (Hills, Maouene, Maouene, Sheya, & Smith, 2009). STAs are not explicitly facts. Rather, they are sufficiently probable properties assumed to be associated with concepts.

Our goal is to determine whether contextualized language models exposed to large corpora encode associations between concepts and their defining properties. We develop probes that specifically test models' ability to recognize STAs.

Probing for Stereotypic Tactic Assumptions

Despite introducing STAs, Prince (1978) provides only a few examples, thus requiring the need to use other data to create probes that evaluate how well contexualized language models capture STAs. We argue that 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,"[1] humans subjects listed properties associated with individual concepts (McRae et al., 2005). When many people individually attribute the same properties to a specific concept, they provide stereotypical tacit assumptions. We use elicited properties that were repeated across human subjects.

Designing Probes We construct prompts for evaluating STAs in LMs by leveraging the CSLB Concept Property Norms (Devereux, Tyler, Geertzen, & Randall, 2013), a large extension of the McRae data, containing 638 concepts each linked with roughly 34 associated properties. The fill-in-the-blank prompts are natural language statements composed of properties where the target concept associated with those human-provided properties is the missing word in the cloze test. If LMs accurately predict the missing concept, we posit that the LMs under consideration encode STAs. 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.[2] Devereux et al. (2013) record production frequencies (PF) enumerating how many people produced each property for a given concept. We select and append the properties with the highest human PF in decreasing order. Iteratively growing prompts enables a *gradient of performance* - we observe concept retrieval based on fewer (clue) properties and track improvements as more are appended.

Probing Method Prompts are fed as input to the neural LM encoder where the t^{th} token is missing. A softmax is taken over the output vector h_t extracted from the model 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 case-sensitive tokens to control for the possibility that a model's vocabulary size influences its rank-based performance.[3] 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 $1/\text{rank}_{\text{LM}}(\text{target concept})$, which is more sensitive to fine-grained differences in rank than just recall, a common retrieval metric. This tracks the predicted rank of a target concept from relatively low ranks given few 'clue' properties to

[1] Instructions shown to participants - specifically appendix B.

[2] 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 (2019) 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.

[3] The vocabulary is the unified intersection of the vocabularies used to train BERT and ROBERTA.

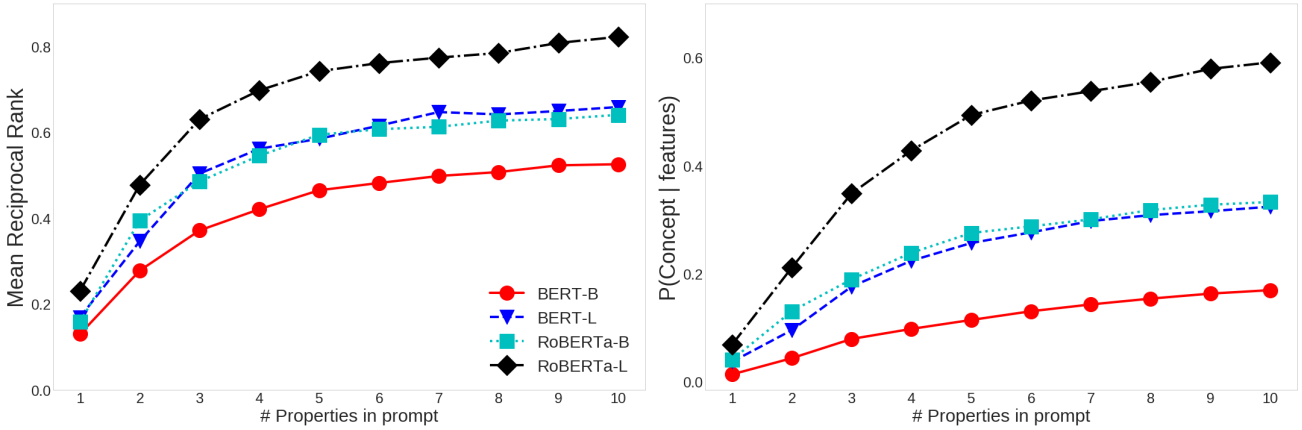

Figure 2: Mean reciprocal rank and probability of correct answer as we increase the number of properties in prompt

much higher ranks as more properties are appended. MRR above 0.5 for a test set indicates that a model top 1 prediction was the correct target concept 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 its empirical performance.

Results

Figure 2 report the results. When given just one property, (x-axis=1) 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. The increase in the MRR and models' confidence (y-axis) as properties are iteratively appended to prompts (increasing x-axis), demonstrates that the LMs more accurately retrieve the correct missing concept when accessing more associated properties. MRR steeply increases for all models as we add more properties to a prompt, but we find less stark improvements beyond after adding four or five properties.

The LARGE models consistently outperform their BASE variants on 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 notably assigns *double* the target probability at 10 properties than that of the next-highest model (RoBERTa-B). Thus, RoBERTa-L is profoundly more confident in its *correct* answers than any other model. However, that all models achieve at least between .5 and .85 MRR conditioned on 10 properties illustrates these contextualized language models' profound ability to identify concepts given their STA sets.

Qualitative analysis We find that model predictions are nearly always grammatical and semantically sensible. RoBERTa-L in particular rarely predicts answers that stray far from the space of the correct answer. 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 later-revealed properties[4]. Not all prompts uniquely identify the target concept, even when a prompt includes 10 properties. [5] However, models still predict answers that are likely to satisfy almost all of the clues.

Categories grouped by properties Are LMs better at retrieving concepts based on different types of properties? We create additional prompts that contain only specific categories. We isolate the CSLB conceptual properties that are grouped into three categories: visual perceptions (bears have fur), functional (bears eat fish), and encyclopaedic (bears are found in forests).[6]

Figure 3a shows that RoBERTa-L performs interchangeably well given encyclopedic or functional type properties alone. In contrast, BERT better retrieves the target concept when given the concept's encyclopedic as opposed to functional properties. Perceptual properties are overall less helpful for models to distinguish concepts compared to non-perceptual properties. This may be the product of category specificity; while perceptual property 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). LMs' ability to retrieve concepts based on associated properties seems to depend based on the type of properties.

Selecting and ordering prompts When designing the probes, we selected and appended the 10 properties with the highest production frequencies (PF) in decreasing PF order. How do these selection and ordering choices affect a models'

[4] 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 the forest'*

[5] e.g. the properties of *buffalo*

[6] We omit properties defined as other perceptions (bears growl) or taxonomic (bears are animals) as few concepts have more than 2-3 such associated properties.

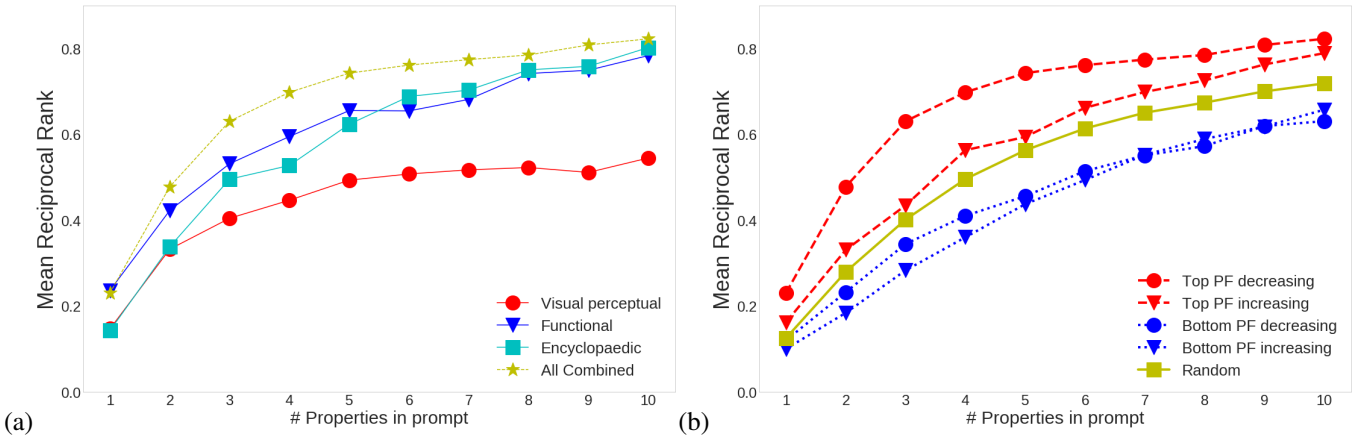

Figure 3: (a) Comparison of RoBERTa-L's performance given only features from each category versus all combined. (b) RoBERTa-L performance on the top-PF versus bottom-PF property sets ordered in increasing vs decreasing PF.

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 property set.[7]

Figure 3b shows the resulting changes 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. These results indicate that the selection and ordering criteria of the properties may matter when adding them to prompts.

Eliciting norming data from language models

We have found that neural language models capture to a surprising degree the relationship between human-produced lists of stereotypic tacit assumptions and their associated concepts. Can we use the LMs to list the properties associated with given concepts under the same type of setup used for human-elicitation? We attempt to replicate the "linguistic filter" (McRae et al., 2005), i.e. lingustic patterns, through which the human subjects convey conceptual knowledge.

In the human elicited studies, subjects were provided "{concept} {relation}..." prompts where the relation could one of four fixed phrases: *is*, *has*, *made of*, and *does*. Subjects were asked to list properties that would fit the prompts. We mimic this protocol using the first three relations:[8] and compare properties predicted by the language models' to the human response set provides.

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

[8]We do not investigate the *does* relation because the resulting human responses are not trivially comparable using template-based prompts. We also construct prompts using *is a* and *has a* for broader coverage of the dataset.

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 are asked directly "What are the properties of X?" (Devereux et al., 2013). This is a highly semantically constraining question that cannot be asked of an off-the-shelf language model. Consequently, 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). It may be unfair to expect language models to replicate how human subjects prefer to more 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. Apprehending this issue, we prepend the phrase 'Everyone knows that' to our new prompts. These prompts therefore take the form of "`Everyone knows that {a bear, a ladder, ...} {is, has, is a, has a, is made of}` ___ ." 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 these contextualized models can only predict a single missing word. This results in an evaluation of between 495 and 583 prompts set for the relations considered.

Feature prediction 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:

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

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 1: Mean average precision and Spearman ρ correlation with human production for LM feature production from concepts. B and R indicate Bert and RoBERTa, b and l indicate base and large models.

where $P_i(j)$ = precision@j and $\Delta r_i(j)$ is the change in recall from item $j-1$ to j for test example i. We report mAP on prediction ranks over a LM's entire vocabulary ($\text{mAP}_{\text{VOCAB}}$), but also re-ranked 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 completions given for a set of concepts are likely to be *wrong* completions for other concepts. While mAP measure the capacity to distinguish the *set*[9] of correct responses from incorrect responses, we also compare probability assigned *within* the set of correct answers by computing average Spearman's ρ between human production frequency and LM probability.

We find that RoBERTa-L outperforms all other versions, sometimes by nearly double mAP. However, we find not insignificant overlap with multiple relations, notably *made of* and *is a*. No model's prediction rank order correlates particularly strongly with that of the human productions frequencies. As discussed below, prompts license completions that are grammatically acceptable but not of the form targeted ('has arrived' as opposed to 'has wheels'). However, when we preclude such completions by narrowing the models' vocabulary to contain only property words, we find that performance (mAP_{SENS}) increases across all models and relations.

Comparing LM probability with human probabilities 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 reflecting how individuals hold different sets of STAs about the same concept. Through either of these lenses, The human subject has produced the sample e.g. 'fur' from some $p(\text{STA} \mid \text{concept} = \textit{bear}, \text{relation} = \textit{has})$ [10]. Like any distribution over language, this can be approximated by a language model and sampled—provided we use appropriate sampling method.

[9] Invariant to order of correct answers

Relation	$\mathbf{mAP}_{\text{VOCAB}}(\Delta)$ Bb	Bl	Rb	Rl
is	-.043	-.031	-.036	*-.101*
is a	*+.113*	+.066	-.001	+.069
has	-.034	-.019	-.092	*-.279*
has a	+.069	+.075	+.029	*-.111*
made of	-.004	+.032	-.052	+.020

Table 2: Change in mean average precision for LM feature production when given prompts with minimized left context

Qualitative analysis of predictions Models generally provide completions that are at least coherent and grammatically acceptable. Most outputs fall at least under the category of 'verifiable of humans,' as McRae note could be listed by humans given sufficient guidance. We also observe properties that apply to the concept but are not reported by humans [11] and properties that apply to senses of a concept that were not considered by the human subjects.[12] We find that some prompts are not sufficently syntactically constraining, licensing non-nominative completions. The pattern *has* permits past participle completions (e.g. 'has arrived') along with the nominative attributes ('has wheels') we target. We do find what could be considered artificial idiosyncracies of models; they favor particular, at times semantically unacceptable relation completions regardless of concept.[13]

Effect of prompt construction on property production To investigate the extent to which our prompt construction encourages property production, we ablate the step in which "everyone knows that" is prepended. Table 2 shows the resulting change in mAP. That changes in prediction accuracy vary so widely by model and relation highlights the difficulty in construct prompt contexts that replicate the 'linguistic lense' through which a LM might produce only concept properties.

[10] 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.

[11] e.g. '*hamsters are real*' and '*motorcycles have horsepower*'

[12] While human subjects list only properties of the *anchor* object concept, the LMs also provide properties that apply to a television anchor.

[13] 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 (0.73), child (0.1), human (0.04), family (0.03), kid (0.02)
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)

Figure 4: ROBERTA-L captures Prince's own exemplary STAs, as shown by assigned probability to both concept and properties.

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 had in mind. Prince lists some of her *own* STAs off the top of her head about the concepts *country* and *person*. We apply the methodologies of the previous experiments and show the resulting conceptual recall and feature productions in Figure 4. 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.

Discussion & Conclusion

We explored the hypothesis owing to Prince (1978) that natural language understanding makes use of types of background knowledge considered stereotypic tacit assumptions. We developed diagnostic experiments derived from human subject responses to a psychological study of conceptual representations and observed that recent contextualized language models trained on large corpora may indeed capture such important information. Through cloze tests, our results provide a lens of quantitative and qualitative exploration of whether BERT and ROBERTA capture concepts and associated properties. We illustrate that the conceptual knowledge elicited from humans by Devereux et al. (2013) is indeed contained within an encoder: that when a human subject may mention something that *'flies'* and *'has rotating blades'*, the LM can infer the description is of a *helicopter*. This may suggest that previous methods for injecting knowledge of semantic features into type-level representations (Fagarasan, Vecchi, & Clark, 2015; Derby, Miller, & Devereux, 2019) may be less necessary for newer contextual encoders. Our work furthers research in probing the extent of semantic knowledge captured by contextualized language models using word prediction tasks.

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 *Proc. of COLING.*

Derby, S., Miller, P., & Devereux, B. (2019). Feature2Vec: Distributional semantic modelling of human property knowledge. *EMNLP*.

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2013). The Centre for Speech, Language and the Brain (CSLB) concept property norms.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.

Ettinger, A. (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models* (Tech. Rep.).

Fagarasan, L., Vecchi, E. M., & Clark, S. (2015). From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *IWCS*.

Goldberg, Y. (2019). Assessing BERT's Syntactic Abilities.

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

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Lucy, L., & Gauthier, J. (2017). Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning. *arXiv:1705.11168 [cs]*. (arXiv: 1705.11168)

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

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.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language Models as Knowledge Bases? *ACL*.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018). Hypothesis Only Baselines in Natural Language Inference. *StarSem*.

Prince, E. F. (1978). On the function of existential presupposition in discourse. In *Papers from the... regional meeting. chicago ling. soc. chicago, ill* (Vol. 14, pp. 362–376).

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

Walker, M. A. (1991). Common Knowledge: A Survey. , 54.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The ieee international conference on computer vision (iccv).*