On the Existence of Tacit Assumptions in Contextualized Language Models

Nathaniel Weir, Adam Poliak, 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 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 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, Naradowsky, et al., 2018), cognitive scientists demonstrate empirically that humans share assumptions about properties associated with concepts (McRae et al., 2005). Taking these conceptual assumptions as one instance of STAs, we ask whether recent contextualized language models trained on large text corpora capture them. In other words, do these 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 given prompts of iteratively concatenated conceptual properties, we find that popular neural language models, e.g. 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). Our analyses indicate that these models associate concepts with perceptual categories of properties (e.g. visual) worse than with non-perceptual ones (e.g. encyclopaedic or functional).

We also 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; Poliak, Haldar, 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. Our diagnostic set and evaluation code will be made publicly available.

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.

| 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, ... |
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. Unidirectional models approximate for any text sequence \( w = [w_1, w_2, \ldots, w_N] \) the factorized distribution \( p(w) = \prod_{i=1}^{N} p(w_i | w_1 \ldots w_{i-1}) \). Recent neural bi-directional language models do not have a well-formed probability expression for entire sequences as they are trained to estimate the probability of an intermediate ‘masked out’ token; this task is colloquially “masked language modelling” (MLM). 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 \( \Pr(w_i | w_1 \ldots w_{i-1}, w_{i+1} \ldots w_N) \). When neural bi-directional LMs that are trained for MLM over large text corpora are subsequently used as contextual encoders,\(^1\) performance across a wide range of natural language understanding tasks drastically improves.

We investigate two recent neural language models: Bi-directional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and Robustly optimized BERT approach (RoBERTa) (Liu et al., 2019). In addition to the MLM objective of predicting randomly-masked words, BERT is trained with an auxiliary objective of next-sentence prediction. BERT is trained on the BookCorpus (Zhu et al., 2015) and English Wikipedia. Using an identical neural architecture, RoBERTa is not trained with the next-sentence auxiliary objective, but is trained on more data with words masked out of larger input sequences. Performance increases ubiquitously on standard NLU datasets when BERT is replaced with RoBERTa as an off-the-shelf contextualized 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 distributions of aggregated human responses in cloze tests designed to target specific behavioral phenomena in sentence processing. 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). 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). 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 of young children (Hills et al., 2009). STAs are not explicitly facts. Rather, they are sufficiently probable properties assumed to be associated with concepts.\(^2\)

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 & Kaai, 2019) have tested for similar types of stereotypic beliefs; they used 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 need to create diagnostics that evaluate how well contextualized language models capture them. 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 stereotypic tacit assumptions. We therefore target the elicited properties that were often repeated across the subjects.

\(^1\)That is, when used to obtain pre-trained contextualized representations of words and sequences.

\(^2\)For example, the STA “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)
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 data that contains 638 concepts each 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 in the cloze test. If LMs accurately predict the missing concept, we posit that they 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.

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 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 with their $r^{th}$ token replaced with a [MASK]. A softmax is taken over the final hidden vector extracted from the model at index $t$ 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. 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 $1/\text{rank}_{LM}(\text{target concept})$, 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 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 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 displays 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 (over the whole vocabulary). The increase in 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 after adding four or five properties. The LARGE models consistently outperform their BASE variants under both MRR and model confidence, as do RoBERTAs over the

---

4Because 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.

5The 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.
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  Examples of prompts and corresponding model predictions are shown in Appendix Table 3. 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 later-revealed properties. Not all prompts uniquely identify the target concept, even when a prompt includes 10 properties. However, models still predict answers that are likely to satisfy almost all of the clues.

Properties Grouped by Category  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 perceptual (bears have fur), functional (bears eat fish), and encyclopaedic (bears are found in forests).8

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). A LM’s ability to retrieve concepts based on associated properties seems to vary 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 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-

---

6E.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’
7E.g. the properties of buffalo do not perfectly distinguish it, as shown in Appendix Table 3.

8We 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 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. The properties with lower PF are correspondingly less beneficial for model performance. This suggests that assumptions about what are less stereotypic—that is, shared among fewer humans—are less well captured by the LMs.

Eliciting Properties 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 protocol defining the “linguistic filter” (McRae et al., 2005), i.e. linguistic patterns, through which the human subjects convey conceptual assumptions.

In the human elicited studies, subjects were provided “{concept} {relation}...” prompts in which the relation could take on one of four fixed phrases: is, has, made of, and does. Subjects were asked to list properties that would complete prompts. We mimic this protocol using the first three relations and compare the properties predicted by the LMs to the corresponding human response sets. Examples of this protocol are shown in Figure 4.

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(STA | concept = bear, relation = has). 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 are asked directly “What are the properties of X?” (Devereux et al., 2014) This is a highly semantically constraining question that cannot be asked of an off-the-shelf language model.

The phrasing of the question also has implications to humans 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 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 mod-

---

The random baseline’s performance is averaged over 5 random permutations of 5 random sets for each concept. 10 Relation was selected at the discretion of the subject via a drop-down menu.

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 also construct prompts using is a and has a for broader coverage of the dataset.

---

| Context | Human Response | PF | ROBERTa-L Response |
|---------|----------------|----|--------------------|
| Everyone knows that a bear has ___ | fur 27 | teeth .36 |
| | claws 15 | claws .18 |
| | teeth 11 | eyes .05 |
| | claws 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 |

Figure 4: Example concept/relation prompts with resulting human and ROBERTa-L responses
We find that some prompts are

\[ \text{mAP} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{|\text{vocabulary}|} P_i(j) \Delta r_i(j) \]

where \( P_i(j) = \text{precision} \) 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 (mAP\textsubscript{VOCAB}), but also re-ranked over a much smaller vocabulary (mAP\textsubscript{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 unlikely completions for other concepts, and that models should be sensitive to this discrepancy. While mAP measure the capacity to distinguish the set\textsuperscript{13} of correct responses from incorrect responses, we also compare probability assigned within the set of correct answers by computing average Spearman’s \( \rho \) between human production frequency and LM probability.

Results using these metrics are displayed in Table 1. 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\textsubscript{SENS}) increases across all models and relations.

Qualitative Analysis 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 et al. note could be listed by humans given sufficient guidance. We observe properties that apply to the concept but are not reported by humans\textsuperscript{14} and properties that apply to senses of a concept that were not considered by the human subjects.\textsuperscript{15} We find that some prompts are not sufficiently 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 idiosyncrasies of models; they favor particular, at times semantically unacceptable relation completions regardless of concept.\textsuperscript{16} We provide examples predictions produced by models in Appendix Table 4.

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

\textsuperscript{13}Invariant to order of correct answers

\textsuperscript{14}e.g. ‘hamsters are real’ and ‘motorcycles have horsepower’

\textsuperscript{15}While human subjects list only properties of the anchor object concept, the LMs also provide properties that apply to a television anchor.

\textsuperscript{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’

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

Table 1: Mean average precision and Spearman \( \rho \) for LM production of properties given concept/relation pairs. B and R indicate BERT and RoBERTA, b and l indicate -BASE and -LARGE models.

| Relation | Data | mAP\textsubscript{VOCAB}(Δ) | Bb  | Bl  | Rb  | Ri  |
|----------|------|-----------------------------|-----|-----|-----|-----|
| is       |      |                             | -.043| -.031| -.036| -.101|
| is a     |      |                             | +.113| +.066| -.001| +.069|
| has      |      |                             | -.034| -.019| -.092| -.279|
| made of  |      |                             | +.069| +.075| +.029| -.111|

Table 2: Change in mean average precision for LM feature production when given prompts with minimized left context
Figure 5: RoBERTA-L captures Prince’s own exemplary STAs (target completions bolded), as shown by predictions of both concept and properties (associated probability in paren).

model and relation highlights the difficulty in constructing prompt contexts that replicate the ‘linguistic lens’ through which a LM might produce only concept 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 5. 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 contextualized 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 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. (2014) 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 et al., 2015; Derby et al., 2019) may be less necessary for newer contextual encoders. We hope that our work serves to further research in exploring the extent of semantic and linguistic knowledge is 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, December). Is an image worth more than a thousand words? on the fine-grain semantic differences between visual and linguistic representations. In Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers (pp. 2807–2817). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from https://www.aclweb.org/anthology/C16-1264

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018, July). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 2126–2136). Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P18-1198 doi: 10.18653/v1/P18-1198

Da, J., & Kasai, J. (2019, November). Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. In Proceedings of the first workshop on commonsense inference in natural language processing (pp. 1–12). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D19-6001 doi: 10.18653/v1/D19-6001

Derby, S., Miller, P., & Devereux, B. (2019, November). Feature2Vec: Distributional semantic modeling of human property knowledge. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-iijcnlp) (pp. 5853–5859). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D19-1595 doi: 10.18653/v1/D19-1595
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), 1119–1127.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers) (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N19-1423
doi: 10.18653/v1/N19-1423

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics, 8, 34–48.

Fagarasan, L., Vecchi, E. M., & Clark, S. (2015, April). From distributional semantics to feature norms: grounding semantic models in human perceptual data. In Proceedings of the 11th international conference on computational semantics (pp. 52–57). London, UK: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W15-0107

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, August). Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In Proceedings of the first workshop on language grounding for robotics (pp. 76–85). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W17-2810
doi: 10.18653/v1/W17-2810

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. Retrieved from https://doi.org/10.1111/j.1467-9280.2007.02028.x (PMID: 18031415) doi: 10.1111/j.1467-9280.2007.02028.x

Poliak, A., Haldar, A., Rudinger, R., Hu, J. E., Pavlick, E., White, A. S., & Van Durme, B. (2018, June). Collecting diverse natural language inference problems for sentence representation evaluation. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 337–340). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-5441
doi: 10.18653/v1/W18-5441

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).

Rubenstein, D., Levi, E., Schwartz, R., & Rappoport, A. (2015, July). How well do distributional models capture different types of semantic knowledge? In Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers) (pp. 726–730). Beijing, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P15-2119
doi: 10.3115/v1/P15-2119
Sommerauer, P., & Fokkens, A. (2018, November). Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 276–286). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-5430 doi: 10.18653/v1/W18-5430

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 International conference on learning representations. Retrieved from https://openreview.net/forum?id=SJzSgnRcKX

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

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018, November). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 353–355). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-5446 doi: 10.18653/v1/W18-5446

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 Proceedings of the ieee international conference on computer vision (pp. 19–27).

Appendix

The following tables show qualitative results of our experiments. Table 3 shows BERT-L and RoBERTA-L’s predicted concepts with associated log probabilities given iteratively longer conjunctions of human-elicited properties. Table 4 shows examples of property production given concept/relation prompts; they are chosen as notable failure cases that exhibit shortcomings of the elicitation and evaluation protocol.
### Table 3: Examples of models’ predicted completions with 1, 5, and 10 ‘clue’ features provided. Associated log probability included in square brackets.

| Context | BERT-L | RoBERTa-L |
|---------|--------|-----------|
| A **bus** has wheels. | car [-2.4], wheel [-2.9], wagon [-3.2], horse [-3.3], vehicle [-3.9] | car [-1.8], bus [-1.9], train [-2.4], bicycle [-2.6], horse [-3.4] |
| A **bus** has wheels, is made of metal, carries a driver, is red, and transports people. | car [-1.6], curt [-2.1], bus [-2.1], truck [-2.7], wagon [-2.9] | bus [-0.6], car [-1.7], train [-2.7], cab [-3.6], taxi [-3.7] |
| A **bus** has wheels, is made of metal, carries a driver, is red, transports people, has seats, is transport, is big, and has windows. | car [-1.1], bus [-1.5], truck [-2.6], vehicle [-3.0], train [-3.2] | bus [-0.8], car [-0.9], train [-3.2], truck [-3.6], vehicle [-3.9] |
| A **cake** is tasty. | bite [-3.1], meal [-3.3], duck [-3.7], little [-3.9], steak [-4.0] | bite [-3.8], steak [-4.1], meal [-4.6], pizza [-4.6], duck [-4.8] |
| A **cake** is tasty, is eaten, is made of sugar, is made of flour, and is made of eggs. | cake [-2.4], dish [-3.2], sweet [-3.6], pie [-3.8], dessert [-3.9] | cake [-2.4], pie [-3.2], cake [-3.4], pie [-2.8], meal [-2.9], banana [-3.4] |
| A **cake** is tasty, is eaten, is made of sugar, is made of flour, is made of eggs, has icing, is baked, is sweet, is a kind of pudding, and is for special occasions. | cake [-1.1], pie [-3.0], dessert [-3.1], jam [-4.0], dish [-4.1] | cake [-1.1], pie [-2.6], cookie [-3.9], dessert [-4.3], cream [-6.5] |
| A **buffalo** has horns. | lion [-2.9], horse [-3.3], goat [-3.6], man [-3.6], bull [-3.9] | bull [-2.8], wolf [-2.9], horse [-3.0], goat [-3.1], cow [-3.3] |
| A **buffalo** has horns, is hairy, is an animal, is big, and eats grass. | goat [-2.6], man [-2.7], horse [-3.1], bear [-3.3], lion [-3.5] | bull [-1.6], cow [-1.8], lion [-2.4], goat [-2.4], horse [-3.1] |
| A **buffalo** has horns, is hairy, is an animal, is big, eats grass, lives in herds, is a mammal, is brown, eats, and has four legs. | man [-1.8], person [-2.0], goat [-2.9], human [-3.3], horse [-3.3] | cow [-1.1], lion [-2.3], bear [-2.5], deer [-2.6], bull [-2.6] |
| A **tiger** has stripes. | number [-4.2], line [-4.2], stripe [-4.3], dot [-4.7], color [-4.8] | tiger [-2.4], dog [-3.4], cat [-3.6], lion [-3.7], bear [-3.7] |
| A **tiger** has stripes, is a cat, is orange, is big, and has teeth. | cat [-1.1], tiger [-2.5], dog [-2.6], person [-3.1], man [-3.6] | tiger [-3.5], cat [-1.9], lion [-2.8], dog [-3.7], bear [-3.7] |
| A **tiger** has stripes, is a cat, is orange, is big, has teeth, is black, is endangered, is a big cat, is an animal, and is a predator. | cat [-4.2], tiger [-2.7], person [-3.5], lion [-4.2], dog [-4.3] | cat [-3.3], tiger [-1.6], lion [-3.5], fox [-4.4], bear [-4.5] |
| A **book** has pages. | page [-0.9], book [-1.2], file [-3.8], chapter [-4.1], word [-4.5] | book [-0.3], diary [-2.2], novel [-2.8], journal [-3.8], notebook [-3.8] |
| A **book** has pages, is made of paper, has a cover, is read, and has words. | book [-0.06], novel [-4.7], manuscript [-4.7], Bible [-5.4], dictionary [-5.5] | book [-0.01], novel [-4.8], newspaper [-6.0], dictionary [-6.7], journal [-7.0] |
| A **book** has pages, is made of paper, has a cover, is read, has words, is found in libraries, is used for pleasure, has pictures, has information, and has a spine. | book [-0.40], note [-4.9], manuscript [-5.4], journal [-5.4], dictionary [-5.9] | book [-0.01], novel [-4.3], dictionary [-6.1], paperback [-6.3], journal [-6.4] |
| A **helicopter** flies. | moth [-1.8], bird [-2.3], fly [-2.7], crow [-3.0], bee [-3.0] | bird [-2.2], bee [-2.4], butterfly [-2.7], bat [-2.9], moth [-3.0] |
| A **helicopter** flies, is made of metal, has rotors, has a pilot, and is noisy. | helicopter [-3.9], bird [-3.1], drone [-3.3], plane [-3.8], rotor [-3.9] | helicopter [-1.2], bird [-4.5], jet [-4.8], airplane [-5.9] |
| A **helicopter** flies, is made of metal, has rotors, has a pilot, is noisy, has blades, has an engine, and carries people. | helicopter [-3.9], plane [-3.0], bird [-3.7], vehicle [-4.2], car [-4.3] | plane [-2.1], helicopter [-2.1], bird [-4.7], airplane [-5.5], aircraft [-5.8] |
| A **taxi** is expensive. | car [-2.5], house [-3.5], divorce [-4.1], ticket [-4.1], house [-4.7] | car [-3.0], house [-4.0], lot [-4.1], life [-4.5], horse [-4.6] |
| A **taxi** is expensive, is yellow, is black, is a cab, and is for transport. | car [-1.0], bicycle [-3.0], vehicle [-3.4], horse [-3.5], bus [-4.1] | Mercedes [-1.8], taxi [-1.9], bus [-2.4], Bentley [-3.0], Jaguar [-3.0] |
| A **taxi** is expensive, is yellow, is black, is a cab, is for transport, is made of metal, has a meter, has wheels, has passengers, and is useful. | car [-2.7], bicycle [-2.3], vehicle [-3.0], horse [-3.8], taxi [-4.2] | taxi [-1.3], bus [-1.5], car [-2.0], bicycle [-2.8], train [-3.5] |
| A **telephone** is made of plastic. | shield [-4.4], chair [-4.4], helmet [-4.5], mask [-4.7], cap [-4.7] | car [-3.1], condom [-3.7], banana [-3.7], toy [-3.8], toilet [-4.0] |
| A **telephone** is made of plastic, is used for communication, has a speaker, rings, and allows you to make calls. | phone [-4.8], telephone [-1.2], mobile [-4.6], receiver [-4.8], cell [-5.1] | phone [-4.6], telephone [-1.3], radio [-5.2], mobile [-6.1], bell [-6.8] |
| A **telephone** is made of plastic, is used for communication, has a speaker, rings, and allows you to make calls, has a receiver, has a wire, is mobile, has buttons, and has a dial. | phone [-4.8], telephone [-1.2], mobile [-4.6], receiver [-4.8], cell [-5.1] | phone [-4.6], telephone [-1.3], radio [-5.2], mobile [-6.1], bell [-6.8] |
| Context                      | Human                     | BERT-B                          | BERT-L                          | RoBERTA-B                     | RoBERTA-L                     |
|------------------------------|---------------------------|---------------------------------|---------------------------------|-------------------------------|-------------------------------|
| Everyone knows that a hamster is ___ .                      | small, alive, cute, white, black  | dangerous, good, right, funny   | dead, real, dangerous, involved | dangerous, evil, bad, dead     | cute, adorable, harmless, alive |
| Everyone knows that a bucket is a ___ .                   | container, vessel, cylinder   | bucket, toilet, problem, mess   | bucket, toilet, weapon, tank    | bucket, toilet, bomb, hat     | toilet, bucket, tool, container |
| Everyone knows that a motorcycle has ___ .                | wheels, seats, lights, brakes, gears | wheels, arrived, escaped, tires | crashed, arrived, died, power   | legs, wings, wheels, power    | wheels, brakes, horsepower, power |
| Everyone knows that an anchor has a ___ .                | chain, cable, rope, point     | problem, story, weakness, camera | purpose, life, weakness, soul   | point, voice, pulse, personality | job, personality, voice, story |
| Everyone knows that a sock is made of ___ .              | cotton, fabric, cloth, material, wool | wood, leather, steel, iron, metal | cotton, rubber, leather, plastic | rubber, wood, metal, plastic, bones | cotton, wool, fabric, rubber, material |

Table 4: Examples of models’ predicted completions to concept/relation prompts targeting the production of properties. Predictions are over the full vocabulary intersection.