Systematic polysemy in adjective-noun combination in contextual word embeddings

Michael Goodale
Ecole Normale Supérieure
michael.goodale@ens.fr

Salvador Mascarenhas
Ecole Normale Supérieure
salvador.mascarenhas@ens.fr

Abstract

In this work, we find evidence that masked language models such as BERT vary their representations of adjectives in a manner that reflects non-trivial theoretical predictions made by linguistic semantic theories about how adjectives compose with nouns. For example, "skilled surgeon" and "skilled carpenter" are skilled in different ways, whereas a "French surgeon" and "French carpenter" are French in the same way. Crucially, we demonstrate via a simple probe that the variation of the adjective depends on the noun in question for adjectives like "skilled" and the noun is not useful for predicting the variation of adjectives like "French". We also show that this variation is systematic—it extends to novel nouns unseen in training. These novel nouns are generated by a process we call "nonce-embeddings", a technique which samples novel embeddings from the manifold of embeddings of nouns in order to generate "meaningless" words akin to nonce-words like *wug* or *blicket* used in linguistics and psychology.

1 Introduction

Natural language is deeply intertwined with context—every utterance is located in a particular linguistic and extra-linguistic setting, and that setting partly determines how to interpret the utterance. While the specific effects of context on language are often rather opaque, this is not always the case. Indeed, many kinds of contextual effects are governed by systematic rules which can be explicitly formulated. Formal semantics as a discipline largely focuses on how words are composed to construct complex meanings, but this task can only be accomplished by modeling the effects of context at various key junctures.

In natural-language processing, contextual word embeddings and other techniques were introduced as a way to handle context at the level of the lexicon, allowing for the same linguistic form to have different meanings in different contexts. The English word "bank" can refer to a feature of fluvial landscape or to a financial institution, and language models must be able to model these two very different meanings.

In linguistic semantics, a crucial distinction is usually made between *polysemy* and *ambiguity*. It is an historically contingent fact that "bank" (the financial institution) and a river "bank" are referred to with the same word—this is *ambiguity* via homonymy for linguistic semanticists. On the other hand, the disambiguated word "bank" which refers to the financial institution participates in a phenomenon of *systematic polysemy*, common to terms related to complex social institutions, where the word can be used both to refer to the institution itself as an abstract social entity and to the building that houses it. Unlike the case of ambiguity via homonymy, these two senses of financial "bank" display *systematicity* and *predictability*, as evidenced by the fact that, say, "mall" or "Department of Motor Vehicles" exemplify the exact same kind of polysemy.

A large and important class of adjectives also display polysemy in this systematic and predictable way. A word like "good" can be seen as having many different meanings depending on the context of utterance; for example, someone who is a "good cook" is someone who is good *at cooking* but possibly terrible at playing an instrument, while the converse is true of a "good musician." These individualized senses of the word "good" are in no way arbitrary, for in general we can tell precisely what the subject of predication is meant to be "good" at by looking at the context. These different meanings for "good" are *derived*, through some combination of compositional and contextual processes, and constitute a case of systematic, and indeed widespread polysemy in the language.

For these and related reasons we review in some detail below, adjectives like "good" have been ar-

gued to have an unpronounced free variable in their structure (e.g. "good-at-*x*"), but the actual representations in people's minds are of course not directly accessible. Unlike people, language models' representations are completely accessible, if hard to interpret. Despite the fact that the representations of words can be looked at directly, the interest in adjective-noun combination has largely been focused on *performance* in both a machine-learning and linguistic sense. That is, previous studies on adjective-noun combination in language models have largely focused on evaluating performance on entailments generated by different kinds of adjectives (Bertolini et al., 2022; Pavlick and Callison-Burch, 2016b,a; Pustejovsky, 2013; Emami et al., 2021).

One of the very few published studies that engage with a version of this question sought to create a compositional model for more traditional static word-embeddings, to see if different kinds of adjectives required different compositional structure, but failed to find a difference between adjective types (Boleda et al., 2013). Yet other work has indicated that transformers may be capable of learning hierarchical compositional structure but did not find evidence that self-supervised language models like BERT do (Murty et al., 2022).

In the present work, rather than trying to manually create a compositional structure on static embeddings for adjectives (Boleda et al., 2013; Baroni et al., 2014), we asked whether language models learn the predicted structure of adjectives and nouns *on their own*.

We report on evidence that language models like BERT do discover, on their own, a rather sophisticated context-sensitive structure highly reminiscent of the theoretical constructs of linguistic semantics. In a nutshell, the models are sensitive to the different ranges of systematic (polysemous) meaning variation that different classes of adjectives allow for. Despite the very minimal theoretical foundation that contextualised word-embeddings emerge from ("words mean different things sometimes"), the model is able to rediscover theoretical distinctions that have been derived explicitly by formal semanticists. These systematic distinctions in adjectives extend even to adjective-noun pairs which are almost certainly outside of the training data and to "nonce-embeddings", a technique we introduce in this work to generate novel semantically void nouns in order to put the systematicity of adjective-noun combinations in these models under stricter scrutiny.

Our results show a partial but striking convergence between theoretical analyses of contextual effects in adjective-noun composition from linguistic semantics and the representations that models like BERT deal in, which of course were learned with no access whatsoever to the linguistic analyses they converge with. This suggests that contextualized word embeddings aren't simply impressive approximators of human *performance*, but may in fact have representations that share sophisticated properties with the representations that have been posited by linguistic semanticists as analyses of the human faculty for language. Additionally, we propose that our novel methodology for nonce-embeddings offers an exciting new way of studying *systematicity* in context dependence by abstracting away less systematic or *un*systematic sources of contextual variation such as frequency of use, morpho-phonological properties, and world knowledge.

2 Methods

2.1 Adjective typology

We defined a simple four-way typology of adjectives based on theories from formal semantics: intersective, weakly subsective, richly subsective and intensional.

Intersective adjectives make a self-contained contribution to the noun phrases they occur in: a "French plumber" is both "French" and a "plumber," and the word "French" makes the same contribution in the NPs "French plumber" and "French CEO."

Weakly subsective adjectives are modulated by the nouns they combine with, which crucially provide the relevant *comparison class*: a "tall five-year old" is likely not "tall" simpliciter, and a "tall basketball player" is tall to a different extent than a "tall five-year old." Semanticists cash out this idea by positing an unpronounced free variable in the structure of adjectives like "tall," determining a (fuzzy) numerical value for height that depends on the noun in question (Kennedy, 2007).

Richly subsective adjectives involve a much more complex free variable: rather than being a numerical value, this free variable denotes an *activity* of some form. A "good plumber" is of course not at all guaranteed to be "good" simpliciter, and a "good CEO" is "good" in a different way than

a "good plumber" is good. Importantly, while it may be tempting to try to analyze "good" the same way as "tall," involving simply a free variable corresponding to a degree of goodness, it is well established that adjectives like "good" necessitate a richer variable. Imagine you and a friend are discussing last night's episode of *Dancing with the stars*, where contestants from all walks of life compete specifically at *dancing*, which is not their primary professional activity. Your friend could say to you "Last night there was an excellent pianist and a rather mediocre actor." The most salient reading of this sentence in this context is that there was a pianist who was excellent *at dancing* and an actor who was mediocre *at dancing*. Such a reading is immediately accounted for by a hidden and rich free variable as we just outlined, which can be filled in either by the linguistic context (the noun it attaches to) or the broader context of utterance (the salient activity of dancing in our example) (Morzycki, 2015).

Finally, *intensional* adjectives so radically combine with nouns that they suspend even entailment to the noun: a "French," "tall," or "good" plumber is at least guaranteed to be a plumber, but an "alleged plumber" is in fact suspected *not* to be a plumber.[1]

2.2 Corpus

To operationalise our typology, we constructed an artificial corpus consisting of sentences where a profession and adjective were combined (e.g. "John is a *good musician*"). We used an artificial corpus to ensure that all adjective noun combinations were evenly distributed and to minimise any difference in embeddings that comes from intervening factors. We constructed a variety of template sentences to account for different syntactic constructions (see Table 2 in the appendix).

Furthermore, the nouns were professions since they are a large class of nouns that we could easily switch adjectives to produce reasonable sentences where no creative interpretation is necessary (i.e. while "a skilled crate" is interpretable, it is very odd as it implies animacy on the part of the crate). While it might be preferable to have a broader semantic category than professions, different adjectives can also change drastically depending on their semantic domain without clear structure. For example, "mythical creature" implies the creature doesn't exist, whereas "mythical status" might simply imply high renown (Pavlick and Callison-Burch, 2016b). By restricting to a small set of interpretable adjectives and nouns, we can be certain that all combinations are coherent and follow our typology consistently even if they may be implausible. Tables 1 and 2 (in the appendix) define schemata for our evaluation corpus and all used words.

2.3 Embedding diversity

A weaker variant of the hypothesis was quickly vindicated by investigating the "self-similarity" of adjective embeddings across different contexts. Self-similarity is just the average cosine similarity of an embedding between different contexts (Ethayarajh, 2019). We found that there was a continuum in how diverse the representations of a given adjective were depending on the noun the adjective combined with. Intersective adjectives varied relatively little, and weak subsective adjectives had a bit more variation, whereas richly subsective and intensional adjectives had considerably more diverse representations with intensionals having the most variation. We tested the following models downloaded from HuggingFace: `bert-base-uncased`, `bert-large-uncased` (Devlin et al., 2019), `sentence-transformers/distilroberta-v1` (Reimers and Gurevych, 2019), `roberta-large` (Liu et al., 2019). All tested models had the reported continuum where intersectives had less diverse representations than weak subsectives, which in turn had less than rich subsectives and with intensionals having the most diverse representations, although different models had different degrees of embedding diversity as shown in previous research (Ethayarajh, 2019).

Crucially, when taking the cosine-similarity across the artificial corpus, we only compared the same adjective used in sentences where everything was identical except for the noun. So, while our corpus incorporated different names, nouns and

[1] This typology can of course be obscured by processes of coercion, as is commonly the case with creative language use. The expression "French pianist" for example might be used to refer not just to someone who is both French and a pianist (its paradigmatic intersective use), but rather a pianist who plays in a distinctively French way, following the French piano school (coercion into a richly subsective reading). We assume in this work that the paradigmatic uses of these adjectives form the bulk of occurrences in the corpora of interest, and that coercion instances between the types we identify aren't so prevalent as to fully nullify the typology. Our results corroborate this assumption.

Adjective Type	Adjectives
Rich	good, bad, skilled, typical, talented, normal, exceptional, terrible, fine, great, horrific, horrible, inferior, dreadful, famous, succesful, unusual, peculiar
Weak	large, fat, nervous, kind, cruel, tall, short, happy, sad, attractive, adventurous, healthy, rich, funny, creepy, foolish
Intersective	bald, straight, naked, gay, white, Black, Canadian, German, Nigerian, Chinese, Brazilian, Christian, Muslim, Jewish, brunette, blond, autistic, diabetic, alcoholic, communist, anarchist, capitalist, socialist
Intensional	alleged, future, potential, presumed, fake, putative, former, occasional, aspiring, failed, amateur, pretend, apparent, wannabe

Table 1: Adjectives sorted according to the typology according to expert annotators (the authors).

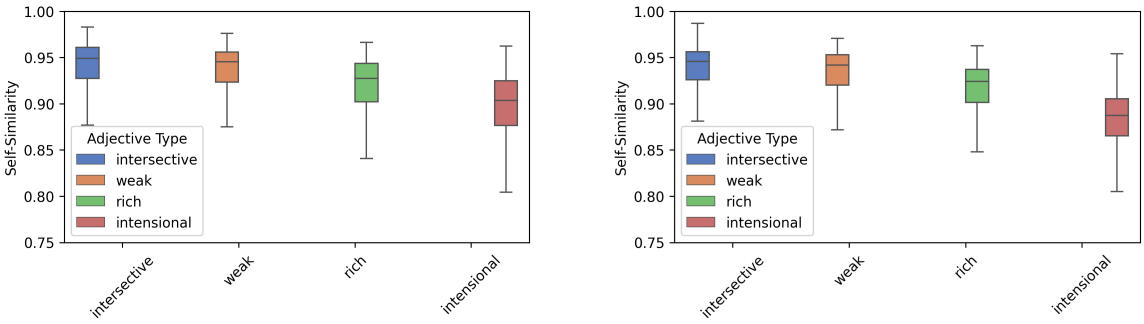
sentence structures, we only took cosine distances between "minimal pairs" of sentences where the only difference was the noun (e.g. "Robert is a skilled surgeon" v.s. "Robert is a skilled carpenter"). As a result, the *only* source of the variation is the different noun used—there is no variation that is attributable to any other intervening factor. We also compared our result to when the minimal pair differed in terms of the name used rather than the noun (e.g. "Robert is a skilled surgeon" v.s. "Phil is a skilled surgeon"). In these contexts, the effect of adjective type disappeared and all adjective types had equal representational diversity reflecting our theoretical predictions.

While it reflects the theoretical distinctions between adjectives, we did not know if BERT or other LLMs derived unique embeddings for adjectives as a function of different nouns, or if BERT simply had various embeddings the same way regular homonymy is handled. In other words, did subsective adjectives have different embeddings in different contexts in the same way that "bank" is ambiguous between meanings if it is in a sentence with financially-related terms or with hydrographic-terms. We investigated this problem in Section 2.4

Intensional nouns We also investigated whether or not nouns also varied depending on the adjective they combined with. Some formal semanticists argue that the meaning of "fake gun" requires a more expansive interpretation of "gun" where "gun" is interpreted to mean more than a regular "gun" (Partee, 2010; Goodale, 2022). The idea is that the standard, literal interpretation of "gun" will contain no fake guns whatsoever (since fake guns aren't guns), so that the expression "fake gun," interpreted literally, would apply to no object in existence. To rescue this defectiveness of "fake gun," the view goes, the noun "gun" has its interpretation *expanded* to include neighboring non-guns, some of which will indeed be "fake guns." If this view is on the right track, we should expect the representation of "gun" to change significantly when combined with an intensional adjective like "fake" but not when combined with an intersective or subsective adjective.

In this situation, the minimal pairs are slightly harder to construct and rely on varying the *adjective* to see what effect that has on the interpretation of the noun. As a result, we represent the data as a confusion matrix across different adjective types (Figure 2). Theoretically, we should expect the nouns that compose with intersectives, weak subsectives and rich subsectives to all be rather uniform in their representations, while the nouns that compose with intensionals should have relatively diverse representations. This is because a "future lawyer" is (currently and actually) a non-lawyer in a very different way than a "fake lawyer" is (currently and actually) a non-lawyer. By contrast, a "French plumber" and a "bald plumber" are plumbers in the same way.

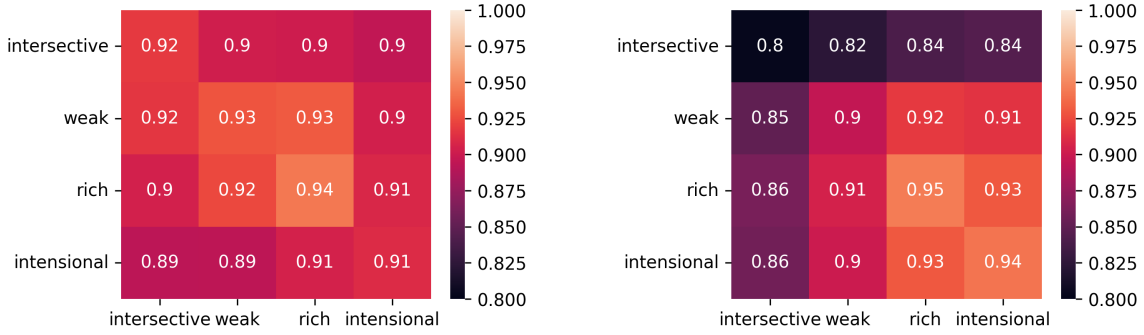
However, our results do not bear out this prediction. For some models, like `bert-large-uncased`, we see very little difference in the representational diversity of nouns across different classes. For others, like `sentence-transformers/distilroberta-v1`, we find that the most diverse representations are among the *intersectives* not the intensionals.



(a) With profession nouns

(b) With generated nonce-embeddings

Figure 1: Self-similarity of adjectives in `bert-large-uncased` across sentences from the evaluation corpus where only the modified noun differs



(a) `bert-large-uncased`

(b) `sentence-transformers/distilroberta-v1`

Figure 2: Confusion matrices showing average cosine similarity for nouns across different adjective types

This is puzzling, because intersective and subsective adjectives (e.g. "French," "good") have non-empty intersections with the nouns they combine with, and thus do not trigger the same rescuing mechanism we described for intensionals (e.g. "fake") above. Thus, we would expect intersective and subsective adjectives to provoke little to no change in the interpretation of the noun they combine with. One interesting possibility is that it could result from the dynamics of typicality. A property such as "Harvard educated" might prompt stereotypes involving wealth and haughtiness, but these stereotypes are not monotonic. For example, experimental subjects have been shown to attribute features to "Harvard-educated carpenters" that they do not attribute to carpenters or to the Harvard-educated, for example "idealistic" (Kunda et al., 1990). This is because the intersection of the Harvard-educated and carpenters will be a population that varies in significant ways from the typical population of either carpenters or the Harvard-educated. Perhaps this puzzling result reflects aspects of this—one would expect a "communist accountant" to be a very different sort of accountant than a "capitalist accountant", even if such a distinction is not predicted by theoretical semantics. Since our results were mixed, we did not investigate the nouns further with probing, but this remains an interesting topic of future research and likely one that relates heavily to problems of fairness in language models (normatively, the representations of "accountant" in "Black accountant" and "white accountant" should not differ significantly).

2.4 Probing

To show that the different embeddings of richly subsective or intensional adjectives were partially *derived* from the modified noun, we used a simple probing technique. By derived, we mean that the meaning of a given adjective is determined by the noun it modifies in a systematic way—the noun does not merely disambiguate between different possible meanings of a noun. This would mean that even completely implausible combinations of adjectives and nouns should display a signature of the same derivational process.

To do this, we created two probes (of a single

linear layer), which were trained to predict a final adjective embedding. One probe was given just the initial embedding of an adjective (including positional embeddings) (AO) whereas the other was given both the initial adjective embedding *and* the noun embedding (AN). Theoretically, we predict that the final representation of an intersective adjective should not be influenced by the noun it combines with, whereas the final representation of a subsective adjective *must* incorporate information from the noun. Unlike intersectives and subsectives, where the view we present in this article is all but universally subscribed to in formal semantics, intensionals are an open question in the literature. On the view we outlined here, they ought to pattern with subsectives. So, we should expect the AO probe to perform roughly as well as the AN probe for intersectives, while for weak subsectives, rich subsectives and intensionals, we should expect the AN probe to steadily get better and better than the AO probe.

To train probes, we did not train directly on our evaluation corpus (because of the small number of adjectives/nouns) but rather on the Simple English Wikipedia (dumped on March 1, 2022 with a CC-BY-SA 3.0 licence). We split the dataset into 90% training and 10% test, with 5751853 sentences total. Sentences were first parsed with SpaCy (`en_core_web_sm`) to tag adjective noun pairs, and then the entire sentence was passed through the language model in question. Afterward, we fed either only the adjective embedding, or the adjective and noun embedding to predict the final adjective embeddings. We only evaluated adjective-noun pairs where the relevant words were a *single* token; we excluded multi-token words. Furthermore, all nouns from our evaluation corpus were excluded from the probe training.

The probes were trained to minimise the cosine distance with the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1e-5 for one epoch. Figure 3 shows the result on the evaluation corpus which we constructed.

3 Systematicity and nonce-embeddings

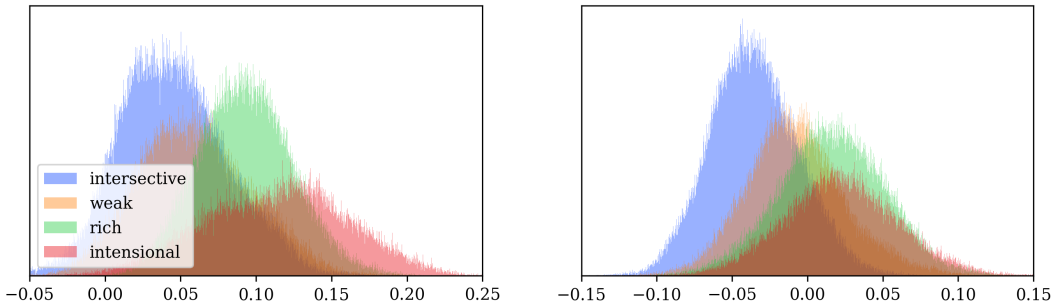
We've demonstrated that the subsective and intensional embeddings are a function of their initial adjective and noun embeddings whereas intersectives rely principally on the adjective. Yet we have not yet characterised how systematic this function is. The probe results alone only indicate that noun information is relevant; it is entirely possible that the different representations of subsective or intensional adjectives arise from the same process for ambiguity we expect for financial banks and river banks. To show that these differences are really *systematic*, we introduce a novel procedure we call nonce-embeddings.

Nonces are meaningless, novel words such as "jabberwocky," "wug," or "blicket" and are commonly used in linguistics and psychology to disambiguate pre-existing lexical knowledge from more general and systematic linguistic abilities. Nonces have been previously used to evaluate GPT-3 (Li et al., 2022), by passing long strings of alphanumeric characters as novel words in a task. The goal was also to evaluate whether the model's behaviour was compositional *and* systematic, but this approach creates novel words comprised of many pre-existing tokens which may put the model at a considerable disadvantage.

Our new technique uses "nonce-embeddings" which allow us to create novel, meaningless words to evaluate the systematicity of models without combining many pre-existing tokens. Rather than passing actual tokens from the model's embedding space, we generate entirely novel, vacuous embeddings that we pass to the model instead of real tokens. We could simply pass random vectors in place of real input token embeddings, but since the input-tokens for models lie on a manifold of some form, this would almost certainly produce bizarre behaviour from the model. Indeed, when we tested with entirely random vectors (random floats between 0 and 1 which are then normalised to a vector magnitude of 1), the effect disappeared completely. Instead, we need vaguely appropriate embeddings just as nonce words for humans must be phonotactically valid.

The technique is simple; we train a generative model on the static, token embeddings from the target language model. In our case, we trained a normalizing-flow model to generate nonce noun embeddings for our target models by training on the set of noun tokens in the input embeddings of a given model. This yields a simple way to generate novel nouns for the model.

Our generative model consisted of 100 RealNVP affine flows (Dinh et al., 2017) combined with ActNorm (Kingma and Dhariwal, 2018) with a hidden size of 64, and nouns were sampled from the input-embeddings of a model if they were a noun in



(a) Evaluation corpus with regular profession nouns

(b) Evaluation corpus with nonce-embeddings

Figure 3: Difference between AN probe's cosine distance from real embedding and the AO probe's distance for `sentence-transformers/all-distilroberta-v1`. A higher value means the noun information was more useful. Note the different scales depending on the type of data.

WordNet. The models were trained with the Adam optimizer with a learning rate of 0.001 for 200 epochs where nouns were sampled proportionally to their frequency in English to make batches of 512 nouns. Generated nonce-embeddings were not simply reproductions of seen nouns; indeed out of 3000 sampled nouns for the `bert-large-uncased` model, the most similar nonce-embedding to any pre-existing embedding had a cosine-similarity of 0.7.

We then evaluated our two probes on 50 generated nonces instead of real profession nouns. The model still showed a clear pattern where the noun information was incorporated for richly subsective and intensional adjectives, but not for intersective adjectives (see Fig 3b). This is evidence that the model is not simply memorising associations of nouns with given adjective embeddings, but is rather *constructing* novel adjective embeddings using the initial noun and adjective embeddings.

4 Results, conclusions, and prospects

To evaluate our model, we measured the cosine similarity between the reconstructed final adjective embedding and the actual adjective embedding produced by the language model (see the appendix for full results). Perhaps unsurprisingly, across all evaluated language models, the adjective-noun (AN) probe performed better than the adjective-only (AO) probe on a validation subset of the training data (Simple English Wikipedia). However, when applied to our evaluation corpus, we found that the AN probe was much more effective than the AO probe for rich subsective and intensional adjectives whereas for the intersective and weak subsective adjectives the AN probe was only marginally better or sometimes worse. This is because noun information is more-or-less irrelevant for intersective adjectives, but it contributes information key to the meaning of the other adjectives, as expected by modern linguistic semantic theories.

Nonce embeddings also preserved the effect for the sentence-transformer model, a sign of systematicity, although for nonces the effect size diminished. The sentence-transformer model had a much larger difference between AO and AN probes for non-intersective adjectives than the other models.

BERT and RoBERTa, did *not* show systematicity with nonces, since the AO probe did better than the AN probe, despite the fact that the embedding-diversity result persisted for BERT, even with nonces (Fig 3b). This means that BERT had different adjective embeddings for different nonces, but the nonce embedding didn't help the probe predict the final adjective embedding. Previous work found evidence for hierarchical compositional behaviour in transformers trained to translate from natural language to an explicit logical semantics, but not in transformers which self-supervise only on natural language (Murty et al., 2022). Sentence-transformer models are trained by supervision to associate paraphrases of sentences and never directly see the semantics. Interestingly, it seems that even the weak signal of paired paraphrases (in the training of the sentence-transformer) may greatly increase the effect of systematic polysemy, absent the heavy-handed semantic annotations of previous approaches.

In sum, we produced evidence of a kind of systematic polysemy in the representations created by large language models which track the theoretical predictions of formal semantics. We also intro-

duced a novel technique for testing the systematicity of large language models: nonce embeddings. This technique could be useful for both evaluating and training different tasks by language models. For example, logical inferences that do not rely on world knowledge (such as those in the HANS dataset, McCoy et al., 2019) could be reinforced by training with a variety of nonce-embeddings rather than exclusively real words.

There are limitations to this study. One, we did not evaluate languages other than English. It would be extremely interesting to extend this analysis to other languages, particularly languages which exhibit unique behaviour related to subsectivity or intensionality. For example, in Russian, most qualitative adjectives have two forms, short and long. The long form can be subsective whereas the short-from will be intersective. For example *studentka umnaja* means a smart-as-a-student student whereas *studentka umna* means a smart-in-a-general-way student. One theoretical analysis of these facts is that in the second case, the adjective has just been blocked from combining with the noun, and the relevant argument is always filled with a broad category like "person" (Siegel, 1976). It would be interesting to see if the representation of *umna* was very close to the representation for *umnaja* when *umnaja* is combined with a semantically broad category like "person" but not when the noun is narrower like "student."

More relevant to the deeper question of polysemy, we did not address multi-token words. This was done largely to keep the analysis as simple as possible, but many nouns (and adjectives) are encoded by language models as sequences of tokens and one should hope that the models preserve this behaviour even if the word is broken up across several tokens. One simple way to extend the model and potentially our results would be to replace our simple linear probe with a very simple recurrent neural network.

Further work should also look at whether this systematic polysemy actually supports a greater performance on NLI tasks involving entailments and subsective adjectives or intensional adjectives. The presence of the kinds of representations we found does not mean they would necessarily be used by models to support entailments, since simpler approximations might be more salient (McCoy et al., 2019). In particular, it is perfectly possible that a model might fail to exhibit the representational properties we described, yet still perform well on an NLI dataset. Indeed, we believe that our results show the importance of complementing performance-oriented criteria with studies such as ours, which aim to understand representations in language models. A model which perfectly models entailment patterns of different adjective types might do so in a manner that is totally different from the kinds of representations that we found in this work.

Indeed, auto-regressive language models likely handle representations in a very different way than the encoder models we investigated, because they do not have the same ability to build representations for specific tokens on the basis on both left and right context. So, while an auto-regressive model might handle adjectives perfectly well from a performance perspective, the representations built for each token will likely not so easily map to theoretical linguistic constructions.

There is one tantalizing reason to believe that masked language models have a deeper connection to formal semantics than auto-regressive models. Formal semantics has a well-established analysis of *focus*, consider the sentence "John gave FLOWERS to Bill," where SMALL CAPS indicate focus, or strong prosodic prominence. The standard view in semantics, dating back to Rooth (1985), is that an open proposition is formed by effectively *masking* the focused constituent: λx.John gave x to Bill. Now this open proposition gets filled in with plausible alternatives, and the propositions thus formed are negated: John didn't give chocolates to Bill, John didn't give apples to Bill, and so on. Focus seems to play a deep and fundamental part in semantics and it is crucially about alternatives generated by an operation eerily reminiscent of masking. The cloze task used by MLM-models might bear a more than superficial relationship to this phenomenon and might lead the language model to representations amenable to contrasting with alternatives, just as humans seem to do. This might mean that while auto-regressive models have all sorts of direct, practical purposes, MLMs may construct representations of language that better reflect theories of linguistic competence.

References

Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in Space: A Program for Composition Distributional Semantics. *Linguistic Issues in*

Language Technology, 9. Publisher: CSLI Publications.

Lorenzo Bertolini, Julie Weeds, and David Weir. 2022. Testing Large Language Models on Compositionality and Inference with Phrase-Level Adjective-Noun Entailment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4084–4100, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Gemma Boleda, Marco Baroni, The Nghia Pham, and Louise McNally. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 35–46, Potsdam, Germany. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2017. Density estimation using Real NVP. ArXiv:1605.08803 [cs, stat].

Ali Emami, Ian Porada, Alexandra Olteanu, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2021. ADEPT: An Adjective-Dependent Plausibility Task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7117–7128, Online. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Michael Goodale. 2022. Manifolds as conceptual representations in formal semantics. Master's thesis, École normale supérieure.

Christopher Kennedy. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. ArXiv:1412.6980 [cs].

Diederik P. Kingma and Prafulla Dhariwal. 2018. Glow: Generative Flow with Invertible 1x1 Convolutions. ArXiv:1807.03039 [cs, stat].

Ziva Kunda, Dale T. Miller, and Theresa Claire. 1990. Combining social concepts: The role of causal reasoning. *Cognitive Science*, 14(4):551–577. Place: Netherlands Publisher: Elsevier Science.

Siyan Li, Riley Carlson, and Christopher Potts. 2022. Systematicity in GPT-3's Interpretation of Novel English Noun Compounds. ArXiv:2210.09492 [cs].

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Marcin Morzycki. 2015. The lexical semantics of adjectives: more than just scales. In *Modification*, Key Topics in Semantics and Pragmatics, pages 13–87. Cambridge University Press.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D. Manning. 2022. Characterizing Intrinsic Compositionality in Transformers with Tree Projections. ArXiv:2211.01288 [cs].

Barbara H. Partee. 2010. 10: Privative Adjectives: Subsective Plus Coercion. In *Presuppositions and Discourse: Essays Offered to Hans Kamp*, pages 273–285. Brill. Section: Presuppositions and Discourse: Essays Offered to Hans Kamp.

Ellie Pavlick and Chris Callison-Burch. 2016a. Most "babies" are "little" and most "problems" are "huge": Compositional Entailment in Adjective-Nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173, Berlin, Germany. Association for Computational Linguistics.

Ellie Pavlick and Chris Callison-Burch. 2016b. So-Called Non-Subsective Adjectives. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 114–119, Berlin, Germany. Association for Computational Linguistics.

James Pustejovsky. 2013. Inference Patterns with Intensional Adjectives. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 85–89, Potsdam, Germany. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Mats Rooth. 1985. *Association with Focus*. Ph.D. thesis, UMass Amherst.

Muffy Siegel. 1976. Capturing the Russian Adjective. In Barbara H Partee, editor, *Montague Grammar*, pages 293–309. Academic Press.

accountant, actor, architect, artist, attorney, banker, bartender, barber, bookkeeper, builder, butcher, carpenter, cashier, chef, coach, dentist, designer, developer, dietician, doctor, economist, editor, electrician, engineer, farmer, filmmaker, jeweler, judge, lawyer, mechanic, musician, nutritionist, nurse, optician, painter, pharmacist, photographer, physician, pilot, plumber, policeman, politician, professor, programmer, psychologist, receptionist, secretary, singer, surgeon, teacher, therapist, translator, translator, undertaker, veterinarian, videographer, waiter, writer

(a) Nouns used to generate the evaluation corpus. Professions were retrieved from a list from Encyclopedia Britannica

James, Mary, Robert, Patricia, John, Jennifer, Michael, Linda, David, Elizabeth, William, Barbara, Richard, Susan, Joseph, Jessica, Thomas, Sarah

(b) Names were taken from the top names in the USA over the last century according to the United States Social Security Administration.

{name} is a {adjective} {noun}.
This is {name}, who is a {adjective} {noun}.
{name} seems like a {adjective} {noun}.
I think {name} is a {adjective} {noun}.
I expect {name} to be a {adjective} {noun}.
If {name} were a {adjective} {noun}, it'd be great.
{name} is known to be a {adjective} {noun}.
A {adjective} {noun} is what {name} is.
{name} might be a {adjective} {noun}.
{name} was a {adjective} {noun}.

(c) Sentence schemata were designed to encompass a broad set of syntactic enviroments.

Table 2: Schemata defining corpus structure

Model name	Probe	Intersective	Weak Subsective	Rich Subsective	Intensional
sentence-transformers/all-distilroberta-v1	AO	$0.725(\sigma = 0.082)$	$0.647(\sigma = 0.063)$	$0.628(\sigma = 0.079)$	$0.577(\sigma = 0.072)$
	AN	$0.771(\sigma = 0.073)$	$0.707(\sigma = 0.062)$	$0.721(\sigma = 0.067)$	$0.691(\sigma = 0.053)$
	AN – AO	$0.045(\sigma = 0.036)$	$0.060(\sigma = 0.037)$	$0.093(\sigma = 0.031)$	$0.113(\sigma = 0.047)$
bert-base-uncased	AO	$0.766(\sigma = 0.053)$	$0.718(\sigma = 0.048)$	$0.715(\sigma = 0.030)$	$0.682(\sigma = 0.042)$
	AN	$0.770(\sigma = 0.047)$	$0.735(\sigma = 0.046)$	$0.739(\sigma = 0.030)$	$0.708(\sigma = 0.041)$
	AN – AO	$0.004(\sigma = 0.018)$	$0.017(\sigma = 0.021)$	$0.024(\sigma = 0.015)$	$0.025(\sigma = 0.023)$
bert-large-uncased	AO	$0.830(\sigma = 0.065)$	$0.803(\sigma = 0.057)$	$0.809(\sigma = 0.035)$	$0.759(\sigma = 0.062)$
	AN	$0.828(\sigma = 0.061)$	$0.805(\sigma = 0.051)$	$0.818(\sigma = 0.034)$	$0.775(\sigma = 0.056)$
	AN – AO	$-0.002(\sigma = 0.013)$	$0.002(\sigma = 0.013)$	$0.008(\sigma = 0.009)$	$0.015(\sigma = 0.015)$
roberta-large	AO	$0.987(\sigma = 0.004)$	$0.986(\sigma = 0.004)$	$0.988(\sigma = 0.004)$	$0.984(\sigma = 0.006)$
	AN	$0.983(\sigma = 0.008)$	$0.981(\sigma = 0.008)$	$0.984(\sigma = 0.007)$	$0.977(\sigma = 0.013)$
	AN – AO	$-0.004(\sigma = 0.007)$	$-0.005(\sigma = 0.007)$	$-0.004(\sigma = 0.006)$	$-0.007(\sigma = 0.012)$

Table 3: Probe performance on evaluation corpus with regular nouns. All reported values are the mean of the cosine distance from the probe's reconstruction and the actual embedding or the difference between the cosine-distances for the two probes (AO–AN).

Model name	Probe	Intersective	Weak Subsective	Rich Subsective	Intensional
sentence-transformer	AO	$0.763(\sigma = 0.074)$	$0.673(\sigma = 0.060)$	$0.671(\sigma = 0.083)$	$0.616(\sigma = 0.060)$
	AN	$0.728(\sigma = 0.072)$	$0.666(\sigma = 0.064)$	$0.688(\sigma = 0.068)$	$0.638(\sigma = 0.056)$
	AN – AO	$-0.035(\sigma = 0.029)$	$-0.008(\sigma = 0.033)$	$0.016(\sigma = 0.035)$	$0.022(\sigma = 0.038)$
bert-base-uncased	AO	$0.740(\sigma = 0.045)$	$0.709(\sigma = 0.047)$	$0.698(\sigma = 0.036)$	$0.666(\sigma = 0.041)$
	AN	$0.700(\sigma = 0.049)$	$0.686(\sigma = 0.057)$	$0.686(\sigma = 0.040)$	$0.636(\sigma = 0.054)$
	AN – AO	$-0.039(\sigma = 0.029)$	$-0.023(\sigma = 0.031)$	$-0.012(\sigma = 0.027)$	$-0.030(\sigma = 0.039)$
bert-large-uncased	AO	$0.813(\sigma = 0.053)$	$0.791(\sigma = 0.055)$	$0.797(\sigma = 0.042)$	$0.740(\sigma = 0.065)$
	AN	$0.778(\sigma = 0.057)$	$0.760(\sigma = 0.057)$	$0.773(\sigma = 0.045)$	$0.715(\sigma = 0.067)$
	AN – AO	$-0.035(\sigma = 0.016)$	$-0.031(\sigma = 0.016)$	$-0.024(\sigma = 0.015)$	$-0.025(\sigma = 0.020)$
roberta-large	AO	$0.988(\pm 0.003)$	$0.987(\pm 0.003)$	$0.988(\pm 0.003)$	$0.984(\pm 0.005)$
	AN	$0.978(\pm 0.008)$	$0.977(\pm 0.008)$	$0.980(\pm 0.007)$	$0.969(\pm 0.013)$
	AN – AO	$-0.010(\pm 0.006)$	$-0.010(\pm 0.007)$	$-0.008(\pm 0.006)$	$-0.015(\pm 0.011)$

Table 4: Probe performance on evaluation corpus with nonce embeddings.