Do contextual word embeddings represent richly subsective adjectives more diversely than intersective adjectives?

Summary. Distributional approaches to adjectival and nominal semantics capture aspects of meaning that have been elusive to mainstream set-theoretic (extensional or intensional) semantics. Yet, *classical* distributional methods have failed to find marks of well-established theoretical classifications from mainstream semantics, in particular the difference between intensional adjectives like ‘fake’ and intersective adjectives like ‘French’ (Boleda et al. 2013). We show that *contextual word embeddings* are sensitive to different kinds of context dependency in the adjectival domain (‘good’ > ‘tall’ > ‘French’). This may demonstrate the systematic distributional differences across the typology of adjectives, and possible evidence for a revision of the typology with respect to intensionals.

Introduction. Traditional set-theoretic approaches to words address important puzzles in compositional semantics and in philosophy of language, but they are do not account well for how meanings of words can change based on what they modify. Privative and subsective adjectives illustrate this clearly, as a skilled mountaineer and skilled violinist are skilled in very different respects. On the other hand, vector-space representations eschew extensions, but make notions of similarity simple. Classical distributional techniques are limited for subsective modification because they are *static*. Models such as BERT (Devlin et al. 2019) however are *contextual*, meaning each representation of a word changes to fit the sentence it is embedded in. These models seem to represent many different linguistic properties such as dependency trees (Hewitt and Manning 2019), and have been successful at syntactic tasks such as subject-verb agreement (Goldberg 2019).

Operationalising the adjective typology. Strictly subsective adjectives change depending on the noun they modify. The meaning of *good* in *good teacher* (i.e. good at teaching) and the meaning of *good* in *good actor* (i.e. good at acting) are radically different. These different meanings are compositional, rather than multiple readings stored in a lexicon.

(1) $[\![\text{Good cook}]\!] = \lambda x_e . [\![\text{good-at-}P]\!](x) \wedge [\![\text{cook}]\!](x)$

Vitally, what a ‘good cook’ is *good at* is left undefined as an extremely rich free variable, standing in for a property or an activity. We call such adjectives RICH SUBSECTIVES. This variable is filled in by the context of the sentence, often from the noun it modifies (e.g. good cooks are good at cooking), but this is not necessarily the case: in (2), the adjectives can be about dancing rather than acting or science.

(2) Last night on Dancing with the Stars, I saw a really talented actor, but also a terrible scientist.

WEAK SUBSECTIVE adjectives take a poorer free variable: only a vague contextual *standard* of comparison (ex: tall). That is, while the meanings of ‘tall’ are plausibly different in ‘tall five-year-old’ and in ‘tall basketball player’, this difference is much weaker than the case of what we call rich subsectives, since the variable to be filled in is much more restricted.

What we call INTERSECTIVES, like “French”, have been historically analyzed without a free variable of any kind. While it is well known that this is an oversimplification (ex: “John is *very* French”), we consider it uncontroversial that the possible variation is much more restrained.

Finally we call non-subsectives, INTENSIONAL adjectives (‘fake’, ‘alleged’).

Contextual word embeddings and adjective modification. Contextual word embeddings have been shown to extend far beyond word-sense disambiguation and have complex semantic and syntactic information (Hewitt and Manning 2019; Tenney et al. 2018). Here we explored the question of whether the subtle variations between different types of adjectives are captured by such models. We expect that the same adjective used in different contexts should have more variation between its contextual embeddings depending on whether it has a free variable and how rich the semantic type of that free variable is.

Method. In order to measure the "semantic divergence" of an adjective, we used the cosine similarity of the embedding of the same adjective in different sentences. We expect richer subsectivity to correspond to greater diversity in representation.

Our dataset was a generated list of simple copular sentences like "NAME is an ADJ NOUN", "NAME is a ADJ NOUN_1 and a NOUN_2", and "NAME is a NOUN_1 and ADJ NOUN_2." This was to control the contexts a word may be used in. We wanted subjects to have as little semantic content as possible, therefore we used proper names. Consequently all nouns were professions.

This then constrained which adjectives we could choose. Specifically, for intersectives, the most widely applicable examples were identity categories ('French', 'Black').

Results and discussion. Figures 1 and 2 show that richer subsectives have a wider diversity in representation than poor subsectives, which in turn vary more than intersectives. This is exactly as predicted, and it shows that distributional data alone is able to capture key differences in adjectival combination. It also tantalisingly suggests that these models might be picking up on cues of contextually filled variables postulated by semantic analyses.

Intensionals were puzzling. We expected intensionals to pattern more closely with intersectives because there is no semantic tradition that postulates rich contextually filled variables for them. Instead, intensionals patterned with subsectives. We note that this class of adjectives is very diverse, with some intensionals being temporal ('former'), modal ('potential'), attitudinal ('alleged'), or none of the above ('fake') and Wikipedia (BERT's training corpus) does not use intensionals often.

This may be connected to BERT's general difficulty with negation (Ettinger 2020), but one intriguing hypothesis is that intensionals *also* have rich hidden variables somewhat like subsectives: 'former' (*when?*), 'possible' (*in what modal scenario?*), 'alleged' (*by whom?*).

We considered if the *noun* may change more with intensionals, but we found that the *noun* self-similarity was not different between intensionals and other adjectives (e.g. no difference between an alleged murderer and a murderer).

It is unclear *how* the representations change in BERT. There are two main possibilities: lexically or compositionally. Lexically implies that the model does not distinguish subsective adjectives from polysemy (for example, the way it distinguishes *bank* for money and river *bank*). Compositionally implies the model does something akin to filling in the free-variable for adjectives. We suspect the model is doing this compositionally as the less polysemous yet subsective adjectives like *skilled* or *talented* have less self-similarity across contexts than polysemous words like *good* or *black*.

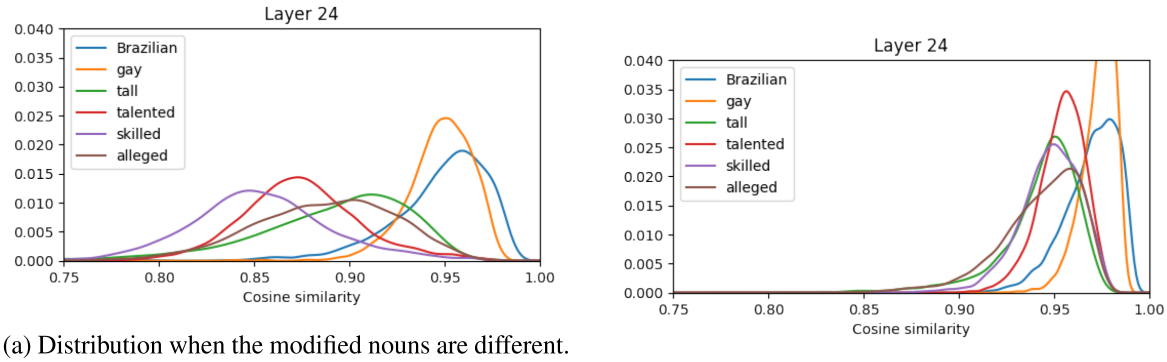
BERT may find a difference between intersectives and subsectives because of the existence of sentences like (3) and the absence of sentences like (4, 5) in corpora.

(3) John is a good musician but not a good friend.

(4) #John is a bald musician but not a bald friend.

(5) #John is a tall musician but not a tall friend.

While BERT has difficulties with negation (Ettinger 2020), it is plausible that it is sensitive enough to the contrast in (3) that it pushes the two contextualised occurrences of 'good' apart, while it gets no such cues for intersectives. Intriguingly, while (5) is possible under certain contexts (e.g. a tall man but not a tall basketball player), we hypothesise similar statements with weak subsectives would be much rarer in corpora. This could account for why BERT only slightly modifies the representations of weak subsectives compared to rich subsectives.



(a) Distribution when the modified nouns are different. Here we expect low self-similarity for subsective adjectives since ‘good’ in ‘good cook’ and ‘good teacher’ are different.

(b) Distribution when the modified nouns are the same. Here self-similarity should be high across the board since the adjective modify the same nouns.

Figure 1: Kernel density estimations of cosine similarities for some adjectives across different sentences. *Brazilian* and *gay* are intersective, *tall* is a weak subsective, *talented* and *skilled* are rich subsectives and *alleged* is intensional.

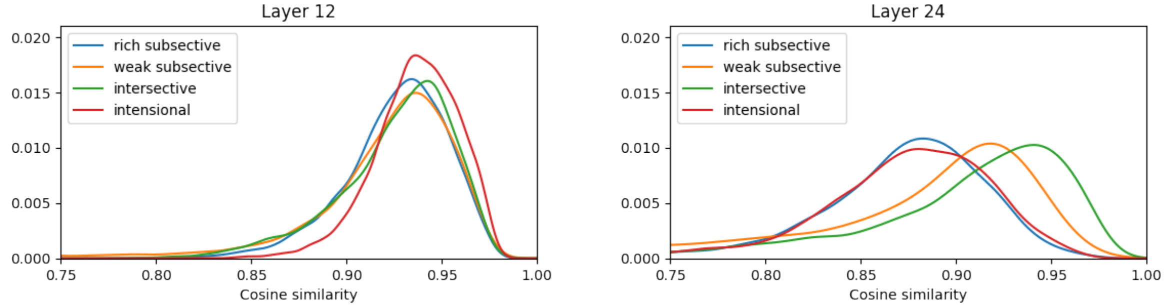


Figure 2: Kernel density estimations of the distribution of cosine similarity for different adjective types when they modify different nouns. The difference between adjective types does not appear until deeper layers (`bert-large-uncased` has 24 layers) where BERT is known to represent more semantic information (Jawahar, Sagot, and Seddah 2019).

Rich Subsective	good, bad, skilled, typical, talented, normal, terrible, fine, great, horrific, horrible, inferior, dreadful
Weak Subsective	large, fat, nervous, kind, cruel, blond, tall, short, happy, sad, beautiful
Intersective	bald, straight, naked, gay, homosexual, white, Black, Canadian, Russian, Chinese, Brazilian
Intensional	alleged, future, potential, presumed

Adjectives

doctor surgeon nurse driver chef waiter scientist plumber gardener carpenter technician athlete actor author novelist musician soldier spy poet

Nouns used in predicate position

I You He She James Mary Robert Patricia John Jennifer Michael Linda William Elizabeth

Names and pronouns used as subjects

[1] Gemma Boleda et al. “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*. Potsdam, Germany: Association for Computational Linguistics, Mar. 2013, pp. 35–46.

[2] Jacob Devlin et al. “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)*. NAACL-HLT 2019. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: `10.18653/v1/N19-1423`.

[3] Allyson Ettinger. “What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models”. In: *Transactions of the Association for Computational Linguistics* 8 (Dec. 2020), pp. 34–48. ISSN: 2307-387X. DOI: `10.1162/tacl_a_00298`.

[4] Yoav Goldberg. “Assessing BERT’s Syntactic Abilities”. In: *arXiv:1901.05287 [cs]* (Jan. 16, 2019). arXiv: `1901.05287`.

[5] John Hewitt and Christopher D. Manning. “A Structural Probe for Finding Syntax in Word Representations”. 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)*. NAACL-HLT 2019. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4129–4138. DOI: `10.18653/v1/N19-1419`.

[6] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. “What Does BERT Learn about the Structure of Language?” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019, pp. 3651–3657. DOI: `10.18653/v1/P19-1356`.

[7] Ian Tenney et al. “What do you learn from context? Probing for sentence structure in contextualized word representations”. In: International Conference on Learning Representations. Sept. 27, 2018.