VAST: The Valence-Assessing Semantics Test for Contextualizing Language Models

Robert Wolfe and Aylin Caliskan
University of Washington
rwolfe3@uw.edu, aylin@uw.edu

Abstract

We introduce VAST, the Valence-Assessing Semantics Test, a novel intrinsic evaluation task for contextualized word embeddings (CWEs). Despite the widespread use of contextualizing language models (LMs), researchers have no intrinsic evaluation task for understanding the semantic quality of CWEs and their unique properties as related to contextualization, the change in the vector representation of a word based on surrounding words; tokenization, the breaking of uncommon words into subcomponents; and LM-specific geometry learned during training. VAST uses valence, the association of a word with pleasantness, to measure the correspondence of word-level LM semantics with widely used human judgments, and examines the effects of contextualization, tokenization, and LM-specific geometry. Because prior research has found that CWEs from OpenAI's 2019 English-language causal LM GPT-2 perform poorly on other intrinsic evaluations, we select GPT-2 as our primary subject, and include results showing that VAST is useful for 7 other LMs, and can be used in 7 languages. GPT-2 results show that the semantics of a word incorporate the semantics of context in layers closer to model output, such that VAST scores diverge between our contextual settings, ranging from Pearson's ρ of .55 to .77 in layer 11. We also show that multiply tokenized words are not semantically encoded until layer 8, where they achieve Pearson's ρ of .46, indicating the presence of an encoding process for multiply tokenized words which differs from that of singly tokenized words, for which ρ is highest in layer 0. We find that a few neurons with values having greater magnitude than the rest mask word-level semantics in GPT-2's top layer, but that word-level semantics can be recovered by nullifying non-semantic principal components: Pearson's ρ in the top layer improves from .32 to .76. Downstream POS tagging and sentence classification experiments indicate that the GPT-2 uses these principal components for non-semantic purposes, such as to represent sentence-level syntax relevant to next-word prediction. After isolating semantics, we show the utility of VAST for understanding LM semantics via improvements over related work on four word similarity tasks, with a score of .50 on SimLex-999, better than the previous best of .45 for GPT-2. Finally, we show that 8 of 10 WEAT bias tests, which compare differences in word embedding associations between groups of words, exhibit more stereotype-congruent biases after isolating semantics, indicating that non-semantic structures in LMs also mask social biases.

Introduction

Contextualizing language models (LMs) are among the most widely used of the ”foundation models” described by Bommasani et al. (2021), a class of powerful but poorly understood AI systems trained on immense amounts of data and used or adapted in many domains. LMs are widely deployed: Google uses BERT for search (Nayak 2019), Facebook uses Linformer for hate speech detection (Schroepfer 2020), and the LMs of the Transformers library of Wolf et al. (2020) are downloaded by millions. However, despite the popularity and use of LMs in consequential applications like medical coding (Salian 2019) and mental health chatbots (Tewari et al. 2021), there is no intrinsic evaluation task - a method to assess quality based on the correspondence of vector geometric properties to human judgments of language - made for contextualized word embeddings (CWEs). Other research assesses CWE semantics using tasks for static word embeddings (SWEs), like SimLex-999, but such tasks are not designed to capture the dynamic behavior of CWEs.

We introduce VAST, the Valence-Assessing Semantics Test, an intrinsic evaluation task for CWEs using valence (association with pleasantness) to measure word-level semantics. VAST is unique among intrinsic evaluation tasks, as it is designed for LMs, and measures LM behavior related to how contextualization (change in the vector representation of a word based on surrounding words), tokenization (breaking of uncommon words into subcomponents), and dominant directions (high-magnitude neurons) affect the semantics of CWEs. VAST takes Pearson's ρ of CWE valence associations and human ratings of valence to quantify the correspondence of CWE semantics with widely held human judgments. We apply VAST to the 12-layer version of the English-language causal LM GPT-2 (Radford et al. 2019). The contributions of VAST are outlined below:

VAST measures the effects of contextualization on word-level semantics. Adaptation to context allows CWEs to differently represent the senses of polysemous words, and CWEs from LMs like GPT encode information about a full sentence (Radford et al. 2018). However, we lack methods for distinguishing when a CWE reflects information related to a word, its context, or both. VAST measures valence in aligned (context has the same valence as the word), misaligned, bleached (no semantic information but the word), and random settings to isolate LM layers where the semantics of the word dominate, and layers where context alters CWEs. GPT-2 VAST scores converge within .01 (ρ) for the random, bleached, and misaligned settings in layer 8, but diverge in the upper layers, with the misaligned setting falling to .55 in layer 11 while the bleached setting stays at .76.

VAST identifies differences in the LM encoding process based on tokenization. Most LMs break infrequent words into subwords to solve the out-of-vocabulary problem. With a large valence lexicon, one can study numerous infrequent words subtokenized by an LM. VAST uses the 13,915-word Warriner lexicon to create large, balanced sets of singly and multiply tokenized words, and isolates where the semantics of multiply tokenized words are comparable to the semantics of singly tokenized words. VAST reveals that multiply tokenized words achieve their highest VAST score in GPT-2 layer 8, at .46, while singly tokenized words begin with a VAST score of .70 in layer 0.

VAST adjusts to dominant directions in CWEs to isolate word-level semantics. Prior research by Mu and Viswanath (2018) found that dominant frequency-related directions distort semantics in SWEs, and that SWEs improve on intrinsic evaluations after nullifying these directions. Other research suggests that the top layers of LMs specialize to their pretraining task (Voita, Sennrich, and Titov 2019), and CWEs from the top layers of causal LMs perform poorly on semantic intrinsic evaluation tasks (Ethayarajh 2019). VAST uses valence to measure the effect of post-processing CWEs using the method of Mu and Viswanath (2018) to isolate word-level semantics. We find that VAST scores fall to .32 in the top layer of GPT-2 in the bleached setting, but that after post-processing, the score improves to .76, with similar improvements for all contextual settings. Moreover, we extract the top directions and use them in experiments which indicate that they encode sentence-level syntactic information useful for next-word prediction.

Drawing on insights from VAST, we outperform GPT-2's scores in prior work on 4 intrinsic evaluations for SWEs. GPT-2 layer 8 CWEs achieve VAST scores of $\rho = .87$ (bleached) and $\rho = .90$ (aligned) with two directions nulled. We use the layer 8 bleached setting CWEs to improve to .66 on WordSim-353 over GPT-2's prior best of .64, and to .50 on SimLex-999 over its best of .45, indicating that VAST also isolates semantic information relevant to word similarity tasks. After isolating semantics, VAST measures social biases, an important step for assessing the potential for harmful associations to manifest in downstream tasks. Isolating semantics allows us to accurately measure CWE associations, and we find that 8 of 10 bias tests on GPT-2 exhibit higher bias effect sizes after doing so.

LMs and Code While an interpretability subfield known as "BERTology" has formed around autoencoders like BERT (Rogers, Kovaleva, and Rumshisky 2021), less research examines CWEs from causal LMs. We apply VAST to GPT-2 because it is the last causal LM made open source by OpenAI, and has scored poorly on intrinsic evaluations despite strong performance on downstream tasks. We also include an appendix of results for 7 LMs and in 7 languages for which we have valence lexica, showing that VAST generalizes. Our code is available at https://github.com/wolferobert3/vast_aaai_2022. We use LMs from the Transformers library (Wolf et al. 2020).

Related Work

We survey related work on word embeddings, evaluation methods for those embeddings, and interpretability research concerning the LMs which produce CWEs.

Word Embeddings SWEs are dense vector representations of words trained on co-occurrence statistics of a text corpus, and have one vector per word (Collobert et al. 2011). SWEs geometrically encode word information related to syntax and semantics, and perform well on word relatedness tasks by measuring angular similarity or performing arithmetic operations on word vectors (Mikolov, Yih, and Zweig 2013). CWEs incorporate information from context (Peters et al. 2018), and are derived from LMs trained on tasks such as causal language modeling (next word prediction), as with the OpenAI GPT LMs (Radford et al. 2018), or masked language modeling (prediction of a hidden "masked" word), as with Google's BERT (Devlin et al. 2019).

Intrinsic and Extrinsic Evaluation Tasks Intrinsic evaluation tasks measure representational quality by how well a word embedding's mathematical properties reflect human judgments of language. Most word similarity tasks measure semantics by taking Spearman's ρ between human ratings and an embedding's cosine similarity for a set of evaluation words (Tsvetkov, Faruqui, and Dyer 2016). Extrinsic evaluation tasks measure performance on downstream tasks, such as sentiment classification (Zhai, Tan, and Choi 2016).

WEAT We measure valence using the Word Embedding Association Test (WEAT) of Caliskan, Bryson, and Narayanan (2017), which evaluates the differential association of two groups of target words related to concepts (*e.g.,* instruments and weapons) with two groups of polar attribute words (*e.g.,* pleasant and unpleasant words). Using groups of attribute words, the WEAT quantifies deterministic biases and differential associations between concepts. The single-category WEAT (SC-WEAT) captures the differential association of one word with two groups of attribute words:

$$\frac{\text{mean}_{a\in A}\cos(\vec{w},\vec{a}) - \text{mean}_{b\in B}\cos(\vec{w},\vec{b})}{\text{std\_dev}_{x\in A\cup B}\cos(\vec{w},\vec{x})} \tag{1}$$

In Equation 1, A and B refer to polar attribute word groups, while $\vec{w}$ refers to a target word vector, and cos refers to cosine similarity. The WEAT and SC-WEAT return an effect size (Cohen's d) indicating strength of association, and a p-value measuring statistical significance. Cohen (1992) defines effect sizes of .2 as small, .5 as medium, and .8 as large. The WEAT uncovered human-like biases, including racial and gender biases, in state-of-the-art SWEs (Caliskan, Bryson, and Narayanan 2017).

The SC-WEAT is similar to lexicon induction methods such as that developed by Hatzivassiloglou and McKeown (1997), in that it can be used to obtain the semantic properties of words without the need for human-labeled data. Lexicon induction methods have been used to infer properties such as the subjectivity, polarity, and orientation of words (Turney and Littman 2003; Riloff and Wiebe 2003).

Valence and ValNorm Valence is the association of a word with pleasantness or unpleasantness, and is the strongest semantic signal hypothesized by Osgood (1964) of valence (referred to by Osgood as evaluation, or as connotation in NLP contexts), dominance (potency), and arousal (activity). Toney-Wails and Caliskan (2021) introduce ValNorm, an intrinsic evaluation task to measure the quality of SWEs based on how well they reflect widely accepted valence norms. ValNorm uses the SC-WEAT to obtain the association of each word in a lexicon with the 25 pleasant words and 25 unpleasant words used in the WEAT. It then takes Pearson's ρ of the SC-WEATs and the human-labeled valence ratings. ValNorm obtains ρ up to .88 on SWEs, and shows that valence norms (but not social biases) are consistent across SWE algorithms, languages, and time periods.

Isotropy Isotropy measures how uniformly dispersed vectors are in embedding space. Anisotropic embeddings have greater angular similarity than isotropic embeddings (Arora et al. 2016). Mu and Viswanath (2018) find that a few dominating directions (vector dimensions) distort semantics in off-the-shelf SWEs like GloVe and Word2Vec, and improve SWE performance on semantic tasks by subtracting the mean vector and nullifying (eliminating the variance caused by) $n/100$ principal components (PCs), where n is dimensionality (Mu and Viswanath 2018).

Transformer Architecture Most LMs adapt the transformer architecture of Vaswani et al. (2017), which uses stacked encoder blocks and decoder blocks for encoding input and producing output. Each block applies self-attention, which informs a word how much information to draw from each word in its context. The GPT LMs are decoder-only causal LMs: they use decoder blocks to produce the next token as output, omitting encoder blocks (Radford et al. 2019). Causal LMs are "unidirectional," and apply attention only to context prior to the input word. LMs are "pretrained" on one or several tasks, such as causal or masked language modeling, which allow LMs to derive general knowledge about language, and then apply that knowledge to other NLP tasks like toxic comment classification in a process known as "fine-tuning" (Howard and Ruder 2018).

Subword Tokenization Most LMs use subword tokenization to solve the out-of-vocabulary problem, which occurs when an LM encounters a word without a corresponding representation. Subword tokenization breaks uncommon words into subcomponents, and ties each subword to a vector in the LM's embedding lookup matrix, which is trained with the LM. LMs form a vocabulary by iteratively adding the subword which occurs most frequently or which most improves the likelihood of the training corpus. This results in two types of word representations: singly tokenized words, which are in the model's vocabulary and can be represented with a single vector; and multiply tokenized words, which are broken into subcomponents, each of which has an associated vector. GPT-2 uses the Byte-Pair Encoding (BPE) algorithm of Sennrich, Haddow, and Birch (2016), which adds the most frequent bigram of symbols to the vocabulary until reaching a defined size. Most intrinsic evaluations use common words to assess semantics, and are ill-adapted to capture the semantic properties of multiply tokenized words.

Representational Evolution Voita, Sennrich, and Titov (2019) find that the layerwise evolution of CWEs in an LM depends on pretraining objective, and show that causal LMs lose information about the current token while predicting the next token. Tenney, Das, and Pavlick (2019) find that BERT approximates an NLP pipeline, with early layers performing well on syntax, middle layers on semantics, and upper layers on sentence-level tasks like coreference resolution.

Bias in LMs LM designers and users must consider worst-case scenarios which might occur as a result using LMs. One of these scenarios, highlighted by Bender et al. (2021) in their work on the limitations of large LMs, involves behavior reflecting human-like social biases that disproportionately affect marginalized groups. Several techniques have been designed for measuring bias in LMs. For example, Guo and Caliskan (2021) treat contextualization in CWEs as a random effect, and derive a combined bias effect size from a meta-analysis of 10,000 WEAT tests. May et al. (2019) insert WEAT target and attribute words into semantically "bleached" templates, such as "This is TERM," to convey little meaning beyond that of the terms inserted to measure bias in sentence vectors from LMs. Sheng et al. (2019) measure "regard" for social groups in LM text output. Nadeem, Bethke, and Reddy (2020) find that LMs with more trainable parameters exhibit better language modeling performance, but prefer biased stereotypes more than smaller LMs. Wolfe and Caliskan (2021) find that under-representation of marginalized groups in the training corpora of four LMs results in CWEs which are more self-similar, but undergo more change in in the model, indicating that LMs generalize poorly to less frequently observed groups, and overfit to often stereotypical pretraining contexts.

Data

VAST requires two data sources: sentences for input to LMs, and lexica with human-rated valence scores.

Reddit Corpus We randomly select one context per word from the Reddit corpus of Baumgartner et al. (2020), which better reflects everyday human speech than the expository language found in sources like Wikipedia.

Valence Lexica VAST measures valence against the human-rated valence scores in Bellezza's lexicon, Affective Norms for English Words (ANEW), and Warriner's lexicon. **The Bellezza Lexicon** of Bellezza, Greenwald, and Banaji (1986) collects 399 words rated by college students on pleasantness from 1 (most unpleasant) to 5 (most pleasant). VAST scores are typically highest with the Bellezza lexicon, which is designed by psychologists to measure valence norms, is the smallest of the lexica, and includes mostly very pleasant or unpleasant words. The ANEW lexicon of Bradley and Lang (1999) includes 1,034 words rated on valence, arousal, and dominance by psychology students. ANEW uses a scale of 1 (most unhappy) to 9 (most happy) for valence. ANEW is commonly used for sentiment analysis. The **Warriner Lexicon** of Warriner, Kuperman, and Brysbaert (2013) extends ANEW to 13,915 words rated on valence, dominance, and arousal by Amazon Mechanical Turk participants.

Word Similarity Tasks We validate VAST by improving on other intrinsic evaluation tasks against scores for CWEs in related work. These tasks use Spearman's ρ between the cosine similarity of each word pair and human-evaluated relatedness. **WordSim-353** (WS-353) consists of 353 word pairs, and was introduced by Finkelstein et al. (2001) to measure information retrieval in search engines, but is widespread as a word relatedness task for SWEs. **SimLex-999** (SL-999) was introduced by Hill, Reichart, and Korhonen (2015) and consists of 666 noun-noun word pairs, 222 verb-verb word pairs, and 111 adjective-adjective word pairs. SimLex evaluates not relatedness but similarity, and has been adapted for multilingual and cross-lingual evaluations by Vulić et al. (2020). **Stanford Rare Words** (RW) labels 2,034 rare word pairs by relatedness, and was designed by Luong, Socher, and Manning (2013) to measure how well a word embedding captures the semantics of uncommon words. Bruni, Tran, and Baroni (2014) introduce the **MEN** Test Collection task, which consists of 3,000 word pairs labeled by relatedness based on responses by Amazon Mechanical Turk participants.

Approach

We provide details related to use of polar words, creation of contextual settings, representation of multiply tokenized words, and PC nullification. The VAST algorithm follows:

1. **Select** a contextual setting (random, bleached, aligned, or misaligned), subword representation (first, last, mean, or max), LM, language, and valence lexicon.
2. **Obtain** a CWE from every layer of the LM for every word in a valence lexicon in the selected contextual setting, using the selected subword representation. If using the misaligned setting, obtain CWEs for polar words in the aligned setting. See the appendix for details about the misaligned setting.
3. **Compute** the SC-WEAT effect size for the CWE from each layer of the LM for every word in the lexicon, using CWEs from the same layer for the polar attribute words in the selected contextual setting. If using the misaligned setting, use the polar word CWEs from the aligned setting.
4. **Take** Pearson's ρ for each layer of SC-WEAT effect sizes vs. valence scores from the lexicon to measure how well LM semantics reflect widely shared human valence norms.
5. **Repeat** the steps above in different contextual settings, using different subword representations, to derive insights about the semantic encoding and contextualization process.

Polar Words VAST measures the strength of the valence signal using the 25 pleasant and unpleasant words from the WEAT, provided in full in the appendix. VAST tokenization experiments use only singly tokenized polar words, as using a set mixed with multiply tokenized words could result in the encoding status of polar words influencing the VAST score of the lexicon words. Most polar words are singly tokenized by LMs; VAST removes all multiply tokenized words from both polar groups, then randomly removes singly tokenized words from the larger group until they are equal in size. For GPT-2, this results in 23 words per polar group.

Contextual Settings To measure whether a CWE reflects the semantics of a word or of its context, we devise four contextual settings: random, semantically bleached, semantically aligned, and semantically misaligned. Where VAST scores diverge between settings, we hypothesize that CWEs are more informed by context. Where VAST scores converge, we observe that CWEs reflect word-level semantics.

In the **Random** setting, each word receives a context chosen at random from the Reddit corpus. In the **Semantically Bleached** setting, each word receives an identical context devoid to the extent possible of semantic information other than the word itself. VAST uses the context "This is WORD", and replaces "WORD" with the target word. In the **Semantically Aligned** setting, each word receives a context reflecting its human-rated valence score.

Templates are matched to words based on human-rated valence scores:
1.0-2.49: It is very unpleasant to think of WORD
2.50-3.99: It is unpleasant to think of WORD
4.00-5.99: It is neither pleasant nor unpleasant to think of WORD
6.00-7.49: It is pleasant to think of WORD
7.50-9.00: It is very pleasant to think of WORD

In the **Semantically Misaligned** setting, each word receives a context clashing with its human-rated valence score. For example, words with the 1.0-2.49 template in the aligned setting are now assigned the 7.5-9.0 template, and vice versa. Words with 4.0-5.99 valence keep their template.

Multiply Tokenized Words Multiply tokenized words are represented by choosing a subtoken vector, or pooling over vectors. We examine four representations: first subtoken, last subtoken, elementwise mean, and elementwise max.

PC Nullification PCs are the "main axes of variance in a dataset," and are sensitive to scale, as obtaining PCs when a few variables have larger magnitude than the rest "recovers the values of these high-magnitude variables" (Lever, Krzywinski, and Altman 2017). Mu and Viswanath (2018) subtract the mean vector and nullify (eliminate the variance caused by) top PCs to restore isotropy in SWEs, which they find improves semantic quality by removing non-semantic information. Ethayarajh (2019) finds that CWEs are anisotropic, and so anisotropic in GPT-2's top layer that any two CWEs have cosine similarity greater than .99. We find that anisotropy in GPT-2's top layer is caused by a few high-magnitude neurons, and apply the method of Mu and Viswanath (2018) to nullify these neurons and restore isotropy. We apply this method to the top and highest scoring layers of GPT-2. While we are primarily interested in uncovering semantics, we also extract the top PCs to study their function in the LM, as described in the next section.

Experiments

VAST experiments measure the effects of contextualization, differences in semantics due to tokenization, and improvements in semantic quality by nullifying top PCs.

Contextualization VAST measures valence against the Bellezza, ANEW, and Warriner lexica using CWEs from every layer of an LM in random, semantically bleached, semantically aligned, and semantically misaligned settings. Where the VAST score is high for all settings, the CWEs reflect the semantics of the current word, rather than its context. VAST scores improving as the layer index increases indicate the semantics of the word being encoded. Divergence of scores between settings points to contextualization.

Tokenization We create large balanced sets of singly and multiply tokenized words. 8,421 Warriner lexicon words are singly tokenized by GPT-2, and 5,494 words are multiply tokenized, allowing for two 5,494-word sets created by randomly selecting a subset of the singly tokenized words. This allows us to examine differences in the semantic evolution of singly and multiply subtokenized words by measuring the layerwise VAST score for each set, and uncover differences in the encoding process for uncommon words. We report results for the random setting.

Semantic Isolation We take an embedding from each word in the Bellezza, ANEW, and Warriner lexica in the semantically bleached setting, and obtain the VAST score before and after applying the PC nullification method of Mu and Viswanath (2018) to remove distorting non-semantic information. This experiment is most relevant to the upper layers of causal LMs, which Voita, Sennrich, and Titov (2019) suggest lose the semantics of the input word while forming predictions about the next word. We apply PC nullification techniques to the top layer and to the highest scoring VAST layer, as even semantically rich SWEs were improved by the method of Mu and Viswanath (2018).

In GPT-2's top layer, 8 neurons make up more than 80% of the length of the 768-dimensional vector. If these neurons solely encode context, the VAST score in bleached and aligned settings should remain high. Our results show that VAST scores fall regardless of setting. To understand what these neurons encode, we extract them using PCA and run two tests with the Corpus of Linguistic Acceptability (CoLA) (Warstadt, Singh, and Bowman 2019). First, we train a logistic regression to predict the acceptability of CoLA test sentences. We take a layer 12 CWE from GPT-2 for the last word in each sentence, and compare performance of the unaltered CWEs, their top PCs, and the CWEs after subtracting the mean and nullifying top PCs. Next we predict the part of speech (POS) of the last word in the sentence using layer 12 CWEs for the last word and next to last word in the sentence, with labels assigned by the NLTK tagger. Power (2020) found that loss for a three-layer neural network trained on POS prediction with layer 12 vectors fell using the next word's POS as a label, rather than the current word's, but this may not be related to top PCs. We first use a binary task, and label classes as Nouns and Not Nouns. Then, a multiclass regression is trained on the NLTK labels. Only "acceptable" CoLA sentences are included.

Validations of VAST While comparison to human judgments is itself a form of validation, we also validate VAST by comparing results on common intrinsic evaluation tasks for the highest scoring VAST layer and the top layer of GPT-2, after semantic isolation in a bleached contextual setting.

Bias VAST uses the WEAT to examine whether human-like biases exist after semantics have been isolated in the top layer of GPT-2 using VAST. WEAT bias tests measure bias on the word level, and use the semantically bleached setting to minimize the influence of context. VAST measures biases using the WEATs introduced by Caliskan, Bryson, and Narayanan (2017). See the appendix for word lists.

Results

Results indicate a middle layer after which semantics of context dominate; a different encoding process based on tokenization; and top PCs which obscure word-level semantics.

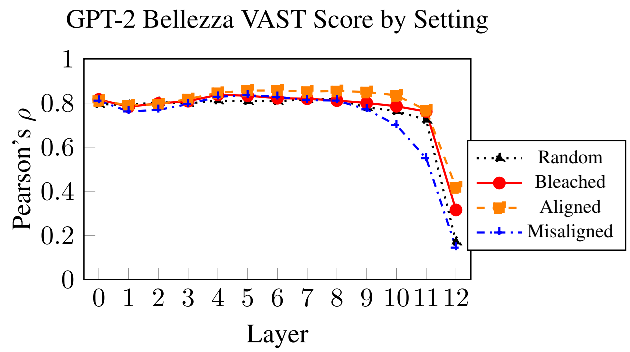

Figure 1: Context alters CWE semantics in the top layers of GPT-2. VAST scores diverge between settings after layer 8.

Contextualization VAST finds that layer 8 of GPT-2 best encodes word-level semantics. ρ in bleached, misaligned, and random settings are within .01 of each other in layer 8 for all three lexica. If representations depended on context, notable differences would exist between settings. Figure 1 shows that differences do arise in the upper layers. The decline by layer 11 is sharpest for the misaligned setting, to .55, followed by the random setting, to .72. We observe broadly consistent results between runs for the random setting, indicating that the natural distribution of contexts in which a word occurs are also reflective of its valence. Scores for aligned and bleached settings stay high, at .77 and .76, and drop in the top layer. Differences among settings reveal a contextualization process for the current word. The same pattern of convergence followed by divergence exists in other causal LMs: for example, in XLNet (Yang et al. 2019), VAST scores among contextual settings are most similar in layer 4, and are the most dissimilar in layer 11 (of 12). On the other hand, settings diverge in autoencoders like RoBERTa from the first layer. Further results are included in the appendix.

Tokenization As seen in Figure 2, VAST reveals that singly tokenized words are semantically encoded in the first layers of GPT-2, with Pearson's ρ on the Warriner lexicon of .70 in the initial layer, but multiply tokenized words are not fully encoded until layer 8, with ρ of .46, revealing a differing encoding process based on tokenization.

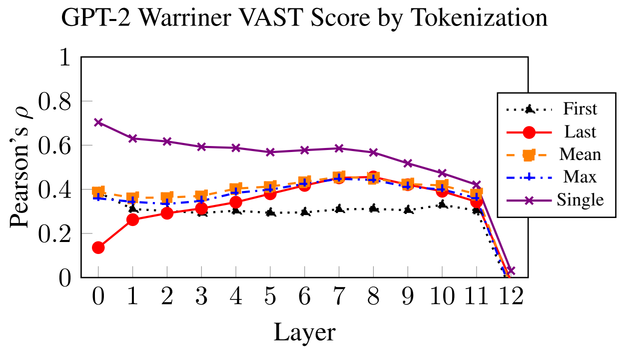

Figure 2: Multiply tokenized words are not encoded until layer 8, but singly tokenized words are encoded in layer 0.

While CWEs from the bottom layers are the least contextualized, and are semantically rich for singly tokenized words, encoding is not complete until later in the LM for multiply tokenized words. Thus, layer 8 CWEs are the most rich for uncommon words, as multiply tokenized words are less frequent in the LM's training corpus. The last subtoken outperforms the mean, indicating that first and middle subtokens contain incomplete semantic information in GPT-2. A similar encoding pattern exists for other causal LMs: in XLNet, the VAST score peaks in layer 5 for multiply tokenized words at .45, and for singly tokenized words in layer 0 at .72. We include further results in the appendix.

Semantic Isolation VAST shows that word-level semantics are changed by context in upper layers of causal LMs, and nearly vanish in the top layer. Table 1 shows results of nullifying top PCs. Word-level semantics are exposed in the top layer after nullifying PCs, but are influenced by context, as the Aligned VAST score improves to .85, but the Misaligned score only to .54. Nullifying 2 PCs in layer 8 improves scores to .87 and .90 for the bleached and aligned settings, showing that non-semantic top PCs exist where the semantic signal is strongest, and that context alters representations even where word-level semantics are most defined. VAST scores improve after nullifying non-semantic PCs in 6 other LM architectures, in every causal LM studied, and in 7 languages in MT5. In XLNet, nullifying 5 PCs improves top layer VAST scores from .56 to .76, with most of the improvement (to .74) coming after nullifying just one PC. Further results are included in the appendix.

VAST (Pearson's ρ) - Semantic Isolation - Bellezza				
Status	Random	Bleached	Aligned	Misaligned
Before	.17	.32	.42	.14
After	.58	.76	.85	.54

Table 1: Nullifying 8 top PCs recovers word-level semantics.

On the CoLA sentence classification task, 11 top PCs from GPT-2 achieve the highest weighted F1 score (.65) of anything except the unaltered top layer CWEs (.65), indicating that top PCs encode information about sentence-level syntax. For POS tagging tasks, the top PCs of the prior word always better predict the POS of the last word than the top PCs of the last word itself. Moreover, 8 PCs of the prior word CWE better predict the POS of the next word than the prior word CWE with those PCs nullified, with an F1 score of .62 for the top PCs and .59 for the nullified CWE. In multiclass POS tagging, the top PCs of the prior word outperform those of the last word, and nearly match the F1 score of the CWE with PCs nullified, which falls to .52 with 12 PCs nullified. This indicates that top PCs in layer 12 encode sentence-level information relevant to predicting the next word.

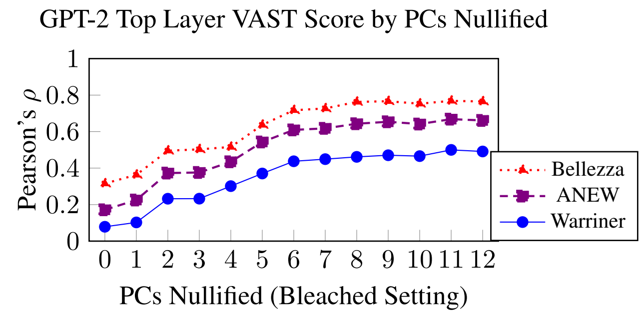

Figure 3: Nullifying non-semantic top PCs exposes word-level semantics in the top layer of GPT-2.

Validations of VAST Figure 4 shows results for GPT-2 layers 8 and 12 on WS-353, SimLex-999, RW, and MEN. CWEs from a semantically bleached context in layer 8 with the mean subtracted and two PCs nulled score .50 on SL-999 and .66 on WS-353, outperforming the previous best GPT-2 results of Bommasani, Davis, and Cardie (2020), who use a pooling method to create an SWE matrix, and results by Ethayarajh (2019), who use the first PC of a word's CWEs.

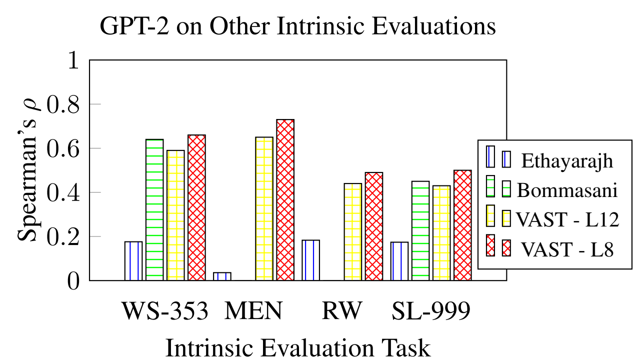

Figure 4: VAST isolates CWEs that outperform related work on other intrinsic evaluations of GPT-2.

Moreover, Toney-Wails and Caliskan (2021) report ValNorm scores as high as ρ = .88 in SWEs. VAST in GPT-2 outperforms this in the aligned setting (ρ = .90), and VAST in the 2.7-billion parameter version of EleutherAI's GPT-Neo (Gao et al. 2020) achieves scores of ρ = .89 (bleached) and .93 (aligned) in layer 12 (of 32), the best results observed on a valence-based intrinsic evaluation task, whether using ValNorm with SWEs or VAST with CWEs.

Bias As shown in Figure 5, isolating semantics with VAST by nullifying non-semantic PCs in GPT-2's top layer exposes both word-level semantics and human-like biases. That bias effect sizes increase as VAST scores improve indicates that the same non-semantic top PCs which distort semantic information in GPT-2 also mask differential biases, which helps to explain why researchers such as Sheng et al. (2019) have found bias in the text output of GPT-2.

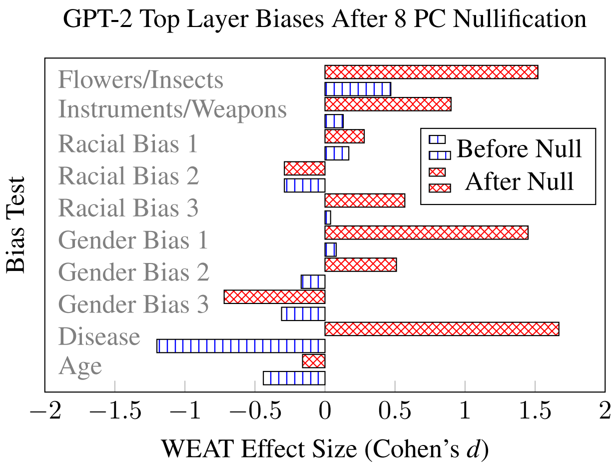

Figure 5: Nullifying dominating directions in the top layer of GPT-2 exposes masked social biases.

Discussion

We suggest 6 reasons to use VAST with CWEs. **First, valence is a well-studied property of language related to NLP tasks like sentiment analysis, whereas it is often unclear what similarity judgments measure.** When VAST scores are low, there is clearly a problem with the embedding's semantics. If an LM associates low-valence words like "murder" and "theft" with the high-valence set (words like "happy" and "joy"), then its associations do not correspond well to human judgments of English semantics. Observing low VAST scores in all settings in layer 12 of GPT-2 led to the insight that the LM's poor performance was not due to the semantics of context, and prompted experiments which found high-magnitude neurons encoding sentence-level information relevant to next-word prediction.

Second, valence can be aligned or misaligned in context to observe the effects of contextualization. Tasks using similarity offer no clear way to create contexts similar or unsimilar to their words, and rate the similarity of word pairs, hindering the creation of such an experiment. VAST can help researchers to determine whether a word-level linguistic association in an LM has been altered by context.

Third, valence is measured for thousands of words, allowing VAST to show that the CWE encoding process differs based on tokenization. Moreover, the lower VAST score of multiply tokenized words (as high as .46 in layer 8, compared to .70 in layer 0 for singly tokenized words) indicates that additional training could benefit these words. Tasks like WS-353 and SL-999 contain more than 90% singly tokenized words for GPT-2, and even though RW has slightly more than half its words multiply tokenized, they are grouped into pairs, resulting in a poorly controlled mix of singly and multiply tokenized word comparisons from which it is hard to draw conclusions about encoding.

Fourth, VAST reveals masked biases in CWEs. High-magnitude neurons in GPT-2's top layer distort word-level semantics, preventing accurate measurement of word associations. Thus, CWE bias measurements may be affected not only by context, which may be controlled for using a meta-analysis method like that of Guo and Caliskan (2021) or a template like that of May et al. (2019), but also by distorting dimensions useful to the LM but problematic for measuring word associations with cosine similarity.

Fifth, VAST is more practical and efficient for measuring CWE semantics than pooling methods. VAST requires just four contexts for each word in a lexicon, resulting in comparatively little compute to obtain useful results.

Finally, VAST can be used in many languages. We apply VAST to 7 languages, with results in the appendix.

Generalization We also apply VAST to the causal LMs XLNet (Yang et al. 2019), GPT-Neo (Gao et al. 2020), and GPT-J, an open-source replication by Wang and Komatsuzaki (2021) of OpenAI's GPT-3 (Brown et al. 2020); to autoencoders BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019); to T5 (Raffel et al. 2020); and to 7 languages (Chinese, Turkish, Polish, French, Spanish, Portuguese, and English) in MT5, a state-of-the-art multilingual adaptation of T5 (Xue et al. 2021). Further analysis is left to future work, and results on these LMs are included in the appendix.

Limitations One limitation of our work is that sentence structure may affect CWEs in ways that distort semantic analysis beyond what VAST detects. Our work suggests the presence of sentence-level information in top PCs, but a more comprehensive study may reveal exactly what these PCs encode. Moreover, while nullifying top PCs reveals stereotype-congruent biases, the effect size for some biases, especially for names, varies with the number of PCs nullified. Further, VAST does not deal with the senses of polysemous words, which would require lexica which label senses by valence. Finally, VAST's tokenization experiment does not control for word frequency, which may affect CWE geometry. We hope to address these limitations in future work.

Conclusion

We introduce VAST, a novel intrinsic evaluation task allowing researchers to measure and isolate word-level semantics in CWEs. VAST reveals how context changes CWE semantics; how tokenization leads to a different encoding process for uncommon words; and how high-magnitude neurons mask word-level semantics - and social biases - in CWEs. VAST allows researchers to extract rich representations from LMs and accurately measure their properties.

References

Arora, S.; Li, Y.; Liang, Y.; Ma, T.; and Risteski, A. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4: 385–399.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, 830–839.

Bellezza, F. S.; Greenwald, A. G.; and Banaji, M. R. 1986. Words high and low in pleasantness as rated by male and female college students. *Behavior Research Methods, Instruments, & Computers*, 18(3): 299–303.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bommasani, R.; Davis, K.; and Cardie, C. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4758–4781.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.

Bradley, M. M.; and Lang, P. J. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Bruni, E.; Tran, N.-K.; and Baroni, M. 2014. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49: 1–47.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Cohen, J. 1992. Statistical power analysis. *Current directions in psychological science*, 1(3): 98–101.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE): 2493–2537.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*.

Ethayarajh, K. 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)*, 55–65.

Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; Wolfman, G.; and Ruppin, E. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, 406–414.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027*.

Guo, W.; and Caliskan, A. 2021. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

Hatzivassiloglou, V.; and McKeown, K. 1997. Predicting the semantic orientation of adjectives. In *35th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics*, 174–181.

Hill, F.; Reichart, R.; and Korhonen, A. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4): 665–695.

Howard, J.; and Ruder, S. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328–339.

Lever, J.; Krzywinski, M.; and Altman, N. 2017. Points of significance: Principal component analysis. *Nature methods*, 14(7): 641–643.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Luong, M.-T.; Socher, R.; and Manning, C. D. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the seventeenth conference on computational natural language learning*, 104–113.

May, C.; Wang, A.; Bordia, S.; Bowman, S.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. 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)*, 622–628.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 746–751.

Mu, J.; and Viswanath, P. 2018. All-but-the-Top: Simple and Effective Postprocessing for Word Representations. In *International Conference on Learning Representations*.

Nadeem, M.; Bethke, A.; and Reddy, S. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nayak, P. 2019. Understanding searches better than ever before.

Osgood, C. E. 1964. Semantic differential technique in the comparative study of cultures 1. *American Anthropologist*, 66(3): 171–200.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Power, A. 2020. Looking for Grammar in all the Right Places.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Riloff, E.; and Wiebe, J. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 105–112.

Rogers, A.; Kovaleva, O.; and Rumshisky, A. 2021. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8: 842–866.

Salian, I. 2019. Cure for the Common Code: San Francisco Startup Uses AI to Automate Medical Coding.

Schroepfer, M. 2020. How AI is getting better at detecting hate speech.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. 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)*, 3407–3412.

Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601.

Tewari, A.; Chhabria, A.; Khalsa, A. S.; Chaudhary, S.; and Kanal, H. 2021. A Survey of Mental Health Chatbots using NLP. *Available at SSRN 3833914*.

Toney-Wails, A.; and Caliskan, A. 2021. ValNorm Quantifies Semantics to Reveal Consistent Valence Biases Across Languages and Over Centuries. *Empirical Methods in Natural Language Processing (EMNLP)*.

Tsvetkov, Y.; Faruqui, M.; and Dyer, C. 2016. Correlation-based Intrinsic Evaluation of Word Vector Representations. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 111–115.

Turney, P. D.; and Littman, M. L. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *acm Transactions on Information Systems (tois)*, 21(4): 315–346.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Voita, E.; Sennrich, R.; and Titov, I. 2019. The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives. 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)*, 4396–4406.

Vulić, I.; Baker, S.; Ponti, E. M.; Petti, U.; Leviant, I.; Wing, K.; Majewska, O.; Bar, E.; Malone, M.; Poibeau, T.; et al. 2020. Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity. *Computational Linguistics*, 46(4): 847–897.

Wang, B.; and Komatsuzaki, A. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Warriner, A. B.; Kuperman, V.; and Brysbaert, M. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4): 1191–1207.

Warstadt, A.; Singh, A.; and Bowman, S. R. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7: 625–641.

Wolf, T.; Chaumond, J.; Debut, L.; Sanh, V.; Delangue, C.; Moi, A.; Cistac, P.; Funtowicz, M.; Davison, J.; Shleifer, S.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Wolfe, R.; and Caliskan, A. 2021. Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 518–532.

Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; and Raffel, C. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 483–498.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhai, M.; Tan, J.; and Choi, J. D. 2016. Intrinsic and Extrinsic Evaluations of Word Embeddings. In *Thirtieth AAAI Conference on Artificial Intelligence*.