Putting Words in BERT's Mouth: Navigating Contextualized Vector Spaces with Pseudowords

Taelin Karidi[1] **Yichu Zhou**[2] **Nathan Schneider**[3] **Omri Abend**[1] **Vivek Srikumar**[2]
[1]Hebrew University of Jerusalem, {taelin.karidi, omri.abend}@mail.huji.ac.il
[2]University of Utah, {flyaway, svivek}@cs.utah.edu
[3]Georgetown University, nathan.schneider@georgetown.edu

Abstract

We present a method for exploring regions around individual points in a contextualized vector space (particularly, *BERT space*), as a way to investigate how these regions correspond to word senses. By inducing a contextualized "pseudoword" as a stand-in for a static embedding in the input layer, and then performing masked prediction of a word in the sentence, we are able to investigate the geometry of the BERT-space in a controlled manner around individual instances. Using our method on a set of carefully constructed sentences targeting ambiguous English words, we find substantial regularity in the contextualized space, with regions that correspond to distinct word senses; but between these regions there are occasionally "sense voids"—regions that do not correspond to any intelligible sense.[1]

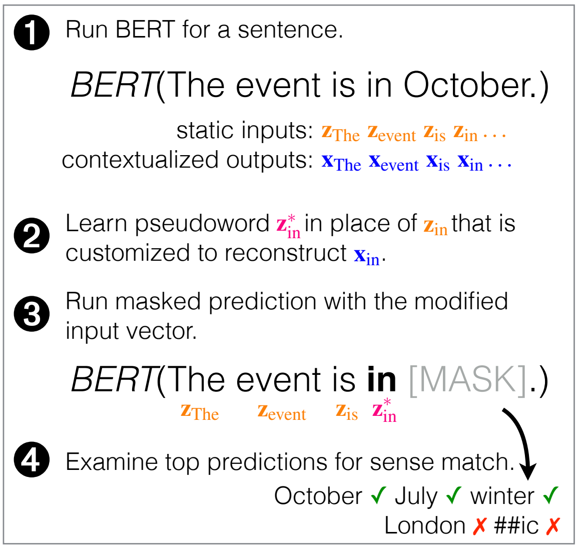
❶ Run BERT for a sentence.
BERT(The event is in October.)
static inputs: $\mathbf{z}_{\text{The}}$ $\mathbf{z}_{\text{event}}$ $\mathbf{z}_{\text{is}}$ $\mathbf{z}_{\text{in}}$ ...
contextualized outputs: $\mathbf{x}_{\text{The}}$ $\mathbf{x}_{\text{event}}$ $\mathbf{x}_{\text{is}}$ $\mathbf{x}_{\text{in}}$ ...
❷ Learn pseudoword $\mathbf{z}^*_{\text{in}}$ in place of $\mathbf{z}_{\text{in}}$ that is customized to reconstruct $\mathbf{x}_{\text{in}}$.
❸ Run masked prediction with the modified input vector.
BERT(The event is **in** [MASK].)
$\mathbf{z}_{\text{The}}$ $\mathbf{z}_{\text{event}}$ $\mathbf{z}_{\text{is}}$ $\mathbf{z}^*_{\text{in}}$
❹ Examine top predictions for sense match.
October ✓ July ✓ winter ✓
London ✗ ##ic ✗

Figure 1: Illustration of the MaPP method as used in the specialization experiments (§5.2). In other experiments, the pseudoword is perturbed prior to step 3.

1 Introduction

Vector spaces defined over static word vectors are somewhat interpretable, as the points are limited to the vocabulary. Contextualized representations (CRs), by contrast, are mysterious because of the unbounded number of distinct contextualized embeddings, and no obvious way to discover the word and context that would correspond to an arbitrary point in the space. Attempts have been made to characterize the information captured in contextualized representations (Rogers et al., 2020; Tenney et al., 2019; Liu et al., 2019), but some of the techniques used (e.g., probing classifiers) have been subject to criticism for their indirectness.

We propose a new technique called **Masked Pseudoword Probing** (MaPP) that allows controlled exploration of the space of a contextualized masked LMs (specifically, English BERT; Devlin et al., 2019). MaPP takes advantage of the static embedding at the first layer of BERT and "hallucinates" new embeddings into this space to correspond to tokens' contextualized representations. By extending BERT's vocabulary with these *pseudowords*, we can use them as inputs for masked prediction of words in the sentence. The words predicted in the masked slot serve as evidence of the meaning of the pseudoword in question—for example, it may encapsulate a specific sense consistent with the original context. We can also transplant a pseudoword into new contexts to see if it generalizes as per our intuitions about word meanings.

We focus on the contextualized meanings of ambiguous verbs and prepositions, which serve as ideal test cases, as they may often be disambiguated by their objects. For example, if a pseudoword interpreted as "in" induces a distribution over its argument slot that gives most of the probability mass to locations such as "London" or "Paris", we may conclude that the pseudoword has a locative sense (table 1). We first ask: How well do pseudowords inferred from contextualized representa-

[1] Our code and dataset are available at https://github.com/tai314159/PWIBM-Putting-Words-in-Bert-s-Mouth

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10300–10313
November 7–11, 2021. ©2021 Association for Computational Linguistics

tions in BERT accord with linguistic expectations about word senses? We investigate this by deploying MaPP for sentences whose masked slot reveals the sense of the ambiguous word. Second, we ask: How semantically smooth is the BERT-space such that arbitrary points near a pseudoword will behave in semantically similar ways? We study this by navigating the space around a pseudoword (or between two pseudowords), and examining the vector's behavior via MaPP. This method allows for investigating the geometry of a contextual representation by traveling the BERT-space in a continuous way and exploring different regions, which we show (see §5) to correspond to distinct concepts.

Our experiments indicate a substantial regularity in the BERT-space. We see regions in the space that correspond to distinct senses. These regions can be recovered using our technique; for example, by sampling points around a pseudoword and looking at the points in the BERT-space which decode to it. Moreover, we see that between sense-regions there are often "voids" in the space that do not correspond to any intelligible sense.

2 Analyzing Contextual Representations

Probing representations. Deciphering the information encoded in contextualized representations like BERT is widely investigated in recent NLP research. *Probing* methods use CRs as inputs to probing classifiers to see how well the CRs may serve as features in predicting specific properties. The intuition is that if the CR can be used to predict a specific property, then knowledge about it is encoded in the representation. Recent classifier-based probes have focused on various linguistic properties such as morphology, parts of speech, sentence length, and syntactic and semantic relations (Liu et al., 2019; Conneau et al., 2018; Belinkov et al., 2017; Adi et al., 2016, *inter alia*). Closely related to ours is work that studied the extent to which the lexical semantic classes of nouns are disambiguated by CRs (Zhao et al., 2020), showing that BERT fares well in this respect.

Beyond classifier-based probes, other approaches have also been explored, such as information theoretic probing (Pimentel et al., 2020; Voita and Titov, 2020), and structural probing (Hewitt and Manning, 2019), which evaluates whether syntax trees are embedded in a linear transformation of a CR's word representation space.

An alternative approach to probing learned representations is directly analyzing the attention weights and activation patterns (Brunner et al., 2020; Abnar and Zuidema, 2020). Criticism against some instances of this approach is found in Jain and Wallace (2019), who claimed that attention weights are less transparent than is often stipulated.

The shortcomings of probing. Some recent work has taken a more critical view regarding probing techniques (Belinkov, 2021). Elazar et al. (2021) argue that while probing methods might show that certain linguistic properties exist in a representation, they do not reveal how and if this information is being used by the probing model. This could be due to the disconnect between the representation itself and the probing model.

Relying on classifiers to interpret representations might be problematic; they add additional confounds to the interpretability of the results, and different representations may need different classifiers (Wu et al., 2020; Zhou and Srikumar, 2021).

Another critique concerns the difference between correlation and causation (Feder et al., 2021): classifier-based probes may rely on shallow correlations in the training set, thus reflecting data artifacts that are irrelevant to the studied distinction.

Word Sense Disambiguation. Word Sense Disambiguation (WSD) aims at making explicit the semantics of a word in context, typically by identifying the most suitable meaning from a predefined sense inventory (Bevilacqua et al., 2021). Disambiguation can also be defined indirectly, through minimal pairs that contrast two senses of a word (Trott and Bergen, 2021) or through another word in the text that determines the semantic class of the word in question (Jiang and Riloff, 2021). Our work bears on this line of work as well: we are using MaPP to test whether the masked prediction indicates that the pseudoword encodes the expected sense. However, we are using carefully controlled sentences so it remains to be seen whether pseudowords can be induced to capture word senses "in the wild".

The geometry of BERT. Understanding the geometry of the BERT-space is not easy. Some attempts in this direction have been made (Coenen et al., 2019; Ethayarajh, 2019; Michael et al., 2020; Mickus et al., 2020; Xypolopoulos et al., 2021; Garí Soler and Apidianaki, 2020), but a more thorough investigation is lacking. As opposed to *predictive methods* such as probing, *descriptive methods* that rely on geometric features of the space analyze the information in CRs directly.

This paper takes a different approach that views BERT as a function that is defined over a continuous space. Our proposed methodology thus allows for a more direct inspection of “gaps” between embedded tokens, that does not require an auxiliary probe and probing dataset, and instead investigates the model’s behavior on arbitrary points in the input space. The paper focuses then on the interpretation of individual points, as opposed to other related work on the geometry of BERT (Coenen et al., 2019; Ethayarajh, 2019; Cai et al., 2021; Hernandez and Andreas, 2021), which mostly considers higher-level properties of the BERT space.

A naïve geometric approach to investigate the information in BERT could be to look at neighborhoods of contextualized embeddings. A vector in this space represents some word within a sentence; it lies in $\mathbb{R}^d$, with $d = 768$. However, it is unclear how such neighborhoods should be defined.

Of course, it is possible to define a discrete neighborhood comprising of contextualized embeddings close to the vector; these may represent the same sense of the same word. Still, in terms of the geometry of the space, how should we interpret a *continuous* neighborhood in the output space? While we could force a non-discrete outlook by generating vectors artificially—e.g., by generating points that are epsilon away from a given point—these artificial contextualized vectors are disembodied, with no obvious linguistic basis. It is therefore unclear how to interpret these artificial vectors (or the linguistic properties of tokens they might encode).

3 Traversing CR Spaces: A New Probing Methodology

Our motivating question is: Are word senses encoded in BERT’s representations—and if so, how? As a test case, we look at highly ambiguous words, as they potentially offer complex geometric configurations of senses in the BERT space.

3.1 Masked Pseudoword Probing

We propose a novel probing technique, *MaPP (Masked Pseudoword Probing)*, which “hallucinates” vectors to reconstruct a token’s contextualized representation. MaPP allows us to “navigate” the BERT-space by looking at neighborhoods of certain word vectors in what we term the *input space*, an extension of the discrete space of BERT’s static (decontextualized) word embeddings. By inducing and manipulating new vectors in the input space (henceforth, **pseudowords**), we can observe the effects on BERT’s behavior via masked prediction. Mathematically, pseudowords are inverse images under the BERT function, continued from the finite space of word embeddings that BERT generally receives, which in the standard implementation contains 30k points in $\mathbb{R}^d$, one per entry in BERT’s vocabulary.[2]

We are interested in the contextualized representation of an ambiguous word, which we call the **focus token** t in a sentence s.

Let $\mathbf{z}_t$ be the static embedding of t (i.e., the input embedding BERT receives for t), and let $\mathbf{x}_t$ be its contextualized representation in s. Let d be the dimension of the input embeddings (without the positional embedding).

To apply our method, we stipulate that there is a specific token called the **cue token**, in the j^{th} position in s, that disambiguates t. Under this assumption, MaPP can discover the sense of a vector $\mathbf{x}$ in the vicinity of $\mathbf{x}_t$, by masking the cue token, and decoding the distribution of its fillers. These fillers serve as a proxy for $\mathbf{x}$’s sense.

MaPP operates in two steps, described next.

1) Pseudoword Representation. First, we find an embedding $\mathbf{z}^*$ that best reconstructs $\mathbf{x}_t$ when input to BERT in place of t. Formally:

$$\mathbf{z}^* = \arg\min_{\mathbf{z}\in\mathbb{R}^d} ||BERT(\mathbf{z}) - \mathbf{x}_t||^2 \qquad (1)$$

where $BERT(\mathbf{z})$ is a forward pass of the model with the vector $\mathbf{z}$ replacing $\mathbf{z}_t$. The solution $\mathbf{z}^*$ is a pseudoword. The original embedding $\mathbf{z}_t$ of t is a solution to eq. (1). But BERT is not invertible, and there is no reason for $\mathbf{z}^*$ and $\mathbf{z}_t$ to be close. Indeed, our experiments show that $\mathbf{z}^*$ is different from t’s input embedding,[3] and suggest that $\mathbf{z}^*$ is a “disambiguated” counterpart of the focus token.

We can approximate $\mathbf{z}^*$ using standard optimization techniques. When solving for $\mathbf{z}^*$, we hold BERT’s parameters fixed, and seek to identify the input embedding $\mathbf{z}$. In standard BERT training, by contrast, the input is known, and we solve for BERT’s parameters.

[2] The term “pseudoword” is also used in psycholinguistics, but refers to a different concept (Gale et al., 1992; Schütze, 1998; Shoemark et al., 2019).

[3] We looked at the distribution of the distances (Euclidean and cosine) between all of the pseudowords and their corresponding static embeddings in the input space.

Focus Word	Sentence A	Sentence B	Sense A	Sense B
in	The event is **in** **October**.	The event is **in** **London**.	temporal	locative
for	The book is **for** **Lisa**.	The book is **for** **reading**.	person	purpose
with	I ate salad **with** **enjoyment**.	I ate salad **with** a **knife**.	feeling	instrument
about	The clip is **about** a **horse**.	The clip is **about** a **minute**.	topic	duration
started	I **started** the **car**.	I **started** the **book**.	device	information source
had	I **had** a **party**.	I **had** a **fever**.	social event	medical condition
had	I **had** **slept**.	I **had** **pizza**.	auxiliary/past participle	food

Table 1: Example minimal pairs from our dataset. Each pair differs in the ambiguous word's argument (and determiner if needed), such that the ambiguous word holds a different sense.

2) Pseudoword-Guided Prediction. After computing $\mathbf{z}^*$, we define a new sentence s', identical to s except for the j^{th} position (the disambiguating position), where we place a mask. For example:

(1) a. s: The event is **in** **London**.
b. s': The event is **in** [MASK].

The focus token t is the ambiguous "in". The cue tokens "London" and "September" would indicate locative and temporal senses of "in" respectively.

Next, we replace the input embedding of t with $\mathbf{z}^*$ or another input space vector $\mathbf{z}$ in its vicinity, and predict the distribution of the slot fillers in the masked position.[4]

3.2 Pseudowords as Input Vectors.

Let us denote the standard input space for the token at t's position (i.e., input embeddings of BERT's vocabulary) with $I_{\text{static}} \subset \mathbb{R}^d$ (where $|I_{\text{static}}| = 30\text{k}$). By extending BERT's inputs to pseudowords, we are performing what is known in mathematical analysis as a *continuation* of BERT's function from the discrete input space $I_{\text{static}} \subset \mathbb{R}^d$ to continuous regions in $\mathbb{R}^d$. This approach allows us to gain insight as to the semantics encoded by different regions of the BERT space. Construing BERT as a continuous function also allows us to invert it, and obtain a point in the inverse image z^* of BERT by solving an optimization problem. We note that viewing the BERT space as a continuous space, e.g., for purposes of mapping between it and other continuous spaces, is an increasingly common practice (Schuster et al., 2019; Gauthier and Levy, 2019); see further discussion in appendix A.4.

In our experiments (§5), the pseudowords will help us explore the geometry of the BERT-space, by traveling across it in a "continuous" way—something that is not possible to do with the BERT vectors as discussed in §2. For example, we can study how perturbations of $\mathbf{z}^*$ (the pseudowords) affect the prediction of the cue word.

4 Experimental Research Questions & The MaPP Dataset

The main hypothesis we study in this paper is:

> There are regular "nicely defined" regions in the BERT-space around words that correspond to distinct senses.

Such regions may be variously interpreted: Around a point with a particular sense, there is a ball which contains mostly points corresponding to that sense. Or, for example, points that correspond to the same sense will lie on a high-dimensional manifold. Cases that we consider as not "nicely defined", are, for example, points corresponding to different senses that are scattered in the space in an inseparable way (or at least inseparable by simple functions). We would like to be able to map the semantic concept of sense to the geometric properties of the BERT-space. However, since little is known about the space, fully characterizing its geometry is out of the scope for this work.

We present MaPP to study this hypothesis, and in doing so introduce the concept of pseudowords. This concept opens additional research questions.

Specialization. Let $\mathbf{z}^* \in \mathbb{R}^d$ be a pseudoword obtained by solving eq. (1) for a sentence s with a focus token t and cue token at position j, holding a sense η. Does $\mathbf{z}^*$ yield a sense distribution (determined by its slot fillers in the j^{th} position) that concentrates on η? That is, does a pseudoword decode to a specific sense of the focus token?

Generalization. Is it possible to transplant a pseudoword into a sentence where the context around the focus token is different, and still obtain coherent results? For example, in the sentences: (a) "The pan is **for** **cooking**." and (b) "The fork is **for**

[4] We also verify that the $\mathbf{z}$ indeed decodes to t, i.e., when $\mathbf{z}^*$ replaces the focus token's embedding, the resulting probability is concentrated on "in".

eating?”, the focus token “for” is the same, and in both cases has a PURPOSE meaning. The context, however, is different. If we transplant the pseudoword for “for” induced with sentence (a), to the position of “for” in a masked version of sentence (b) (i.e, “The pan is **for** [MASK]?”), will we get a coherent prediction with the same sense? Or is the pseudoword obtained from one sentence limited to a specific context?

If pseudowords obtained for one sentence do not generalize to another, we propose to induce a “generalized pseudoword” trained over multiple examples with the same sense.

4.1 The MaPP Dataset

To answer the questions listed above, we manually compiled the MaPP Dataset, a controlled dataset with short sentences, designed to avoid confounds that may introduce difficulties in interpreting the results.

Each sentence contains an ambiguous word that is fully disambiguated by a specific slot in the sentence. E.g., in the sentence “The book is **for reading**”, the ambiguous word “for” has a PURPOSE sense, strongly signaled by “reading”. All sentences were reviewed by a linguist to maximize naturalness and minimize ambiguity.

The dataset consists of 3 portions, each used in different experiments. We describe each portion adjacent to the relevant experiment.

Relational words as a test case. We chose to focus our analysis on the ambiguity of relational words in English, specifically prepositions and verbs. Relational words present an interesting test case: many are highly ambiguous and encode basic semantic distinctions, such as space, time and manner (Schneider et al., 2018). We do not attempt to cover all possible senses of the selected words; instead, we have constructed our dataset to illustrate just a few clear contrasts (see further discussion in appendix A.4).

5 Experiments

We use MaPP to empirically evaluate the hypotheses listed above.

We conduct four types of experiments. First, we test specialization, the extent to which the induced pseudoword $\mathbf{z}^*$ can be viewed as a sense-disambiguated version of the focus token’s input embedding. Second, we venture into the immediate regions around $\mathbf{z}^*$ by perturbing it, and examining how this affects the resulting senses. We thereby gain insight as to the regularity of the region around $\mathbf{z}^*$ with respect to the focus token’s sense. Third, we examine the sense regularity of the BERT CRs by examining the senses encountered when traversing the line between two pseudowords corresponding to different senses (e.g., the locative and temporal senses of “in”). Finally, we test generalization, namely, the extent to which $\mathbf{z}^*$ can serve as a sense-disambiguated embedding of the focus word *in different contexts*.

Query		Top 5 predictions
The dinner is **on Monday**.	$\mathbf{z}$	fire ✗ offer ✗ sale ✗ Friday ✓ hold ✗
	$\mathbf{z}^*$	Sunday ✓ Saturday ✓ Thursday ✓ Tuesday ✓ Friday ✓
The clip is **about** a **queen**.	$\mathbf{z}$	minute ✗ year ✗ second ✗ day ✗ week ✗
	$\mathbf{z}^*$	woman ✓ girl ✓ man ✓ child ✓ boy ✓

Table 2: Specialization examples where the pseudoword $\mathbf{z}^*$ learned from the query sentence corresponds to a different sense from BERT’s static word embedding $\mathbf{z}$, as evidenced by the top 5 predictions when the cue token (**Monday**, **queen**) is masked out.

5.1 Implementation Details

Throughout, we use the BERT-base-cased model via the implementation in HuggingFace (Wolf et al., 2020) and Pytorch (Paszke et al., 2019).

To solve for $\mathbf{z}^*$ (eq. (1)), we add a new token to the vocabulary (`#TOKEN#`), which corresponds to the *focus token*. When backpropagating the gradients, we ensure that the gradients of all parameters of BERT are zero, except the token embedding of `#TOKEN#`. In this way, we preserve the original BERT model while enabling us to solve for $\mathbf{z}^*$. We use 5 random initializations and select the z^* with the lowest loss. We use standard gradient-based optimization for this process. We are solving for the input to BERT rather than model parameters, so we backpropagate through the network, holding BERT’s parameters fixed, and take gradients with respect to z.

5.2 Specialization Experiments

We test whether we can control the sense of BERT’s predicted tokens using a *pseudoword* $\mathbf{z}^*$. If this is indeed the case, it supports our view of $\mathbf{z}^*$ as a disambiguated version of its corresponding static token embedding. In this experiment, we designate highly ambiguous words—specifically, verbs like “have” and prepositions like “in”—as the focus tokens and apply the process described in §3.1 on these ambiguous words.

SENSE MATCH	All		Verbs		Prepositions	
	@1	**@5**	**@1**	**@5**	**@1**	**@5**
N (total # of predictions)	94	470	43	215	51	255
Vanilla BERT (**z**)	39.8	36.0	36.4	27.9	41.0	42.7
MaPP ($\mathbf{z}^*$)	77.6	65.1	83.7	67.0	72.5	63.5

Table 3: Specialization results: accuracy at producing a completion consistent with the sense from the original context (out of top-1 and top-5 predictions). Subscores are provided for verb & preposition focus words.

WORD MATCH	All			Verbs			Prepositions		
	@1	**@5**	**@20**	**@1**	**@5**	**@20**	**@1**	**@5**	**@20**
N	94	470	1880	43	215	860	51	255	1020
Vanilla BERT	5.3	17.0	35.1	2.3	11.6	20.9	7.8	21.6	47.1
MaPP	24.5	51.1	74.5	25.6	55.8	79.1	23.5	47.1	70.6

Table 4: Specialization results: rate of predicting the word that was masked in the original sentence (recall in top k predictions).

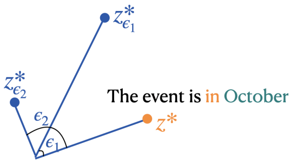

Figure 2: Perturbation of $\mathbf{z}^*$ by different ε values.

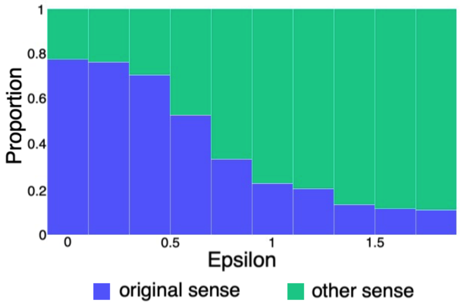

Figure 3: ε-perturbation: average accuracy at producing a completion consistent with the sense from the original context, for each ε (average over 10 directions per ε). On average, as ε increases, the rate of matching the original sense decreases.

Data. We use the **Basic Portion** of the MaPP Dataset for this experiment. It contains 94 sentences with 8 ambiguous words: *had*, *started*, *run*, *in*, *for*, *with*, *about*, and *on*. Two to four senses of each word type are represented with 5 sentences each (except for rare cases where we could not find five coherent sentences for a certain template).

Evaluation. For each sentence we perform masked prediction and manually categorize which of the top 5 predicted words are consistent with the original sense. From these judgments we compute the accuracy—the proportion of sense-congruent predictions.[5] In total, we evaluate 470 predictions.

Results. Table 3 shows the performance of MaPP versus a Vanilla BERT baseline (using the static embeddings rather than pseudowords for masked prediction). We see that in most cases, by applying MaPP, we shift the prediction of the model to the desired sense, which establishes the validity of our technique. Further, table 4 shows that typically, after applying MaPP, the model's top prediction is not the word that was masked in the original sentence—i.e., the pseudoword is not simply memorizing a specific cue word. Table 2 illustrates two examples exhibiting a clear shift to the desired sense. Not every pseudoword behaves as expected, though, as is discussed in subsequent experiments.

5.3 ε-Perturbation

Our goal in this experiment is to "travel" in the BERT-space. Since it is not clear how to interpret a direct perturbation of the contextualized vectors, we do this via the input space. We compute a pseudoword and perturb it to obtain new points in an ε-ball around it, as schematized in figure 2.

Given a pseudoword $\mathbf{z}^*$ for a particular token in context, normalize it to a unit vector. Choose 10 random directions w by sampling uniformly from the unit sphere. For each direction w and perturbation distance ε, find a vector w' that is ε away (in cosine distance) from $\mathbf{z}^*$ in the direction w (i.e., w' is on the intersection between the plane spanned by $\mathbf{z}^*$ and w, and the unit sphere). We do this for several values of ε (see below). Each perturbed $\mathbf{z}^*$ is fed back into the model and used for masked prediction.

Data. In this experiment we use the **Basic Portion** of the MaPP Dataset, as described in §5.2.

Evaluation. For $\varepsilon \in \{0, 0.2, ..., 1.8\}$ and each direction, we examine at the model's top 5 predictions for the masked token to determine which are consistent with the original sense. We measure the average accuracy over the 10 directions for each ε.

Results. The fraction of predicted words consistent with the original sense decreases gradually as the amount of perturbation ε increases (figure 3 and numeric results in appendix A.4). This matches our hypothesis that there is regularity in the BERT space, and that it is carved into regions which correspond to distinct senses. Outside of these regions (where ε is large), we occasionally encounter *sense voids*—regions where there is no intelligible sense

[5] We also conducted a random baseline experiment, with randomly sampled vectors from $\mathbb{R}^d$ instead of the pseudoword. The accuracy was negligible.

Mask	Vanilla BERT	Query 1	Interpolated MaPP				Query 2
			$\alpha = 0$	$\alpha = 0.4$	$\alpha = 0.8$	$\alpha = 1$	
The event is in [MASK].	progress ✗	The event is **in London**.	**London** ◀	Toronto ◀	June ▶	July ▶	The event is **in August**.
	June ▶		Dublin ◀	**London** ◀	July ▶	September ▶	
	July ▶		Edinburgh ◀	June ◀	March ▶	June ▶	
	April ▶		Paris ◀	Dublin ◀	September ▶	March ▶	
	September ▶		Sydney ◀	Melbourne ◀	April ▶	**August** ▶	
The book is for [MASK].	children ◀	The book is **for him**.	me ◀	children ◀	free ✗	free ✗	The book is **for viewing**.
	women ◀		her ◀	women ◀	sale ▶	download ▶	
	adults ◀		**him** ◀	you ◀	download ▶	sale ▶	
	sale ▶		you ◀	sale ▶	reading ▶	reading ▶	
	boys ◀		us ◀	free ✗	children ◀	purchase ▶	
I started the [MASK].	engine ◀	I **started** the **car**.	engine ◀	engine ◀	engine ◀	**book** ▶	I **started** the **book**.
	car ◀		**car** ◀	**car** ◀	**car** ◀	story ▶	
	motor ◀		ignition ◀	ignition ◀	**book** ▶	game ▶	
	truck ◀		truck ◀	truck ◀	machine ◀	engine ◀	
	ignition ◀		motor ◀	motor ◀	fire ✗	movie ▶	
I had [MASK].	to ✗	I **had cake**.	anxiety ✗	it ✗	to ✗	won ▶	I **had gone**.
	it ✗		it ✗	doubts ✗	it ✗	died ▶	
	nothing ✗		worry ✗	to ✗	been ▶	left ▶	
	him ✗		energy ✗	problems ✗	not ✗	**gone** ▶	
	her ✗		power ✗	anxiety ✗	won ▶	forgotten ▶	
The clip is about a [MASK].	**minute** ◀	The clip is **about** a **minute**.	**minute** ◀	**minute** ◀	woman ▶	woman ▶	The clip is **about** a **queen**.
	year ◀		second ◀	second ◀	man ▶	girl ▶	
	second ◀		third ✗	third ✗	**minute** ◀	man ▶	
	day ◀		minutes ✗	minutes ✗	day ◀	child ▶	
	week ◀		moment ◀	moment ◀	year ◀	boy ▶	
The dinner is on the [MASK].	table ◀	The dinner is **on** the **plate**.	table ◀	same ✗	evening ✗	**Wednesday** ▶	The dinner is **on Wednesday**.
	rocks ◀		house ✗	usual ✗	weekend ▶	Sunday ▶	
	way ✗		menu ✗	winner ✗	Sunday ▶	weekend ▶	
	beach ◀		line ✗	evening ✗	**Wednesday** ▶	same ✗	
	menu ✗		way ✗	weekend ▶	village ✗	Saturday ▶	

Table 5: Example top-5 predictions with interpolated MaPP versus Vanilla BERT. ◀ indicates that the prediction has been coded as consistent with the Query 1 sense, ▶ for the Query 2 sense, and ✗ for neither. Predictions that result in an ungrammatical sentence are also coded as ✗. The expectation is that values of α closer to 0 will be more reflective of the Query 1 sense, while values closer to 1 will be more reflective of the Query 2 sense. (Note that 0.4 and 0.8 are not evenly spaced between 0 and 1.) Original words from the queries are bolded.

compatible with the context. For example, with the query "The event is **in Canada**.", we see small values of ε producing names of countries, but $\varepsilon \geq 0.6$ producing adjectives ("annual", "amateur", "contested", "open", "free") which are ungrammatical and nonsensical in context.

5.4 Interpolation

Next, we take two pseudowords representing distinct senses of an ambiguous word in minimal pair sentences, and traverse the space between them to determine what the boundary between sense regions looks like. Given two pseudowords $\mathbf{z}_1$ and $\mathbf{z}_2$, we simply interpolate their vectors: $\mathbf{z}^*_\alpha = (1-\alpha)\mathbf{z}^*_1 + \alpha\mathbf{z}^*_2$, where $0 \leq \alpha \leq 1$ controls how much weight to put on one pseudoword or the other. This is depicted in figure 4.

Data. In this experiment we use the **Minimal Pairs Portion** of our dataset. These 40 pairs of sentences differ only in the cue to give contrasting senses of the focus word.[6] 7 different ambiguous words and 16 distinct senses appear in this portion of the dataset, with 5 sentences for each distinct sense. Several examples appear in table 1.

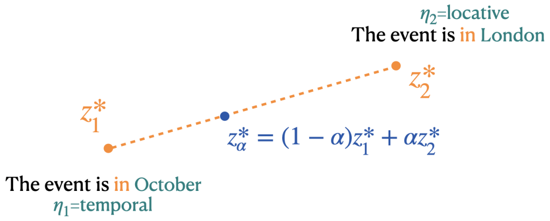

Figure 4: Illustration of the interpolation process.

Evaluation. For each sentence in a minimal pair, we infer **z**. Then for $\alpha \in \{0, 0.1, 0.15, 0.2, \ldots, 1\}$, we compute $\mathbf{z}^*_\alpha$, use it for masked prediction, and judge whether each of the top 5 predictions corresponds to the sense in the first sentence, the second sentence, or neither.

Results. Figure 5 shows the overall proportion of predictions for each sense as α progresses from 0 to 1. We see a gradual trend from one sense to the other. This matches our hypothesis that there is regularity in the BERT-space: traveling on a line between two senses in the input spaces decodes to two distinct regions in the BERT-space.

For some individual examples, there is a sharp boundary at some α; for others there is an intermediate region where the predictions mix the two senses or are unrelated to both (see appendix A.4).

The behavior with static embeddings ("Vanilla BERT") can serve as a control for interpreting the effect of the interpolated pseudoword, as shown for

[6] In some cases, the two elements of a minimal pair differ syntactically as well. Viz.: auxiliary vs. main verb ("I **had gone**/**cake**"); verb-particle construction vs. verb+PP ("run **over** the **cat**/**bridge**"); and PP vs. approximation modifier ("**about** a **horse**/**minute**").

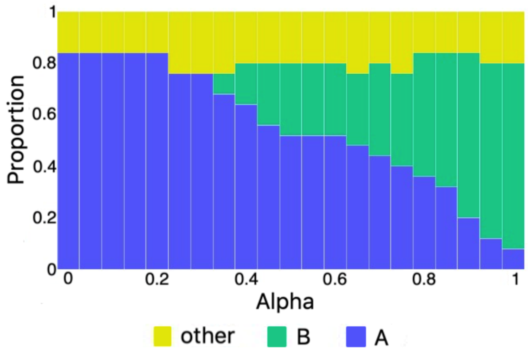

Figure 5: Interpolation results for minimal pair data as a function of interpolation parameter α: average proportion of top-1 predictions consistent with sense A, which predominates at $\alpha = 0$; sense B, which predominates at $\alpha = 1$; or neither.

several examples in table 5. In many cases Vanilla BERT prefers one of the two senses by default, but with α sufficiently close to the other sense's pseudoword, the behavior changes. In general, the transition from one sense to the other is readily apparent from the predictions. Exceptions where one of the expected senses is inadequately represented include "I **had** **cake**." (no foods are predicted in the top 5) and "The dinner is **on** the **plate**." (not as many food-oriented locations were predicted as expected).

5.5 Generalization Experiments

In this experiment we examine whether the pseudoword is specialized for a particular sense of the focus word only in a context-specific fashion, or whether the pseudoword is a valid representation of the sense in new contexts where the focus word may appear. To this end we take a pseudoword from one context and "transplant" it into a new context. We are particularly interested in transplantations where the ambiguous word has a similar meaning but is expected to yield a new distribution of masked predictions, due to the influence of the new context. For example:

(2) a. s: The book is **for** **reading**.
b. s': The cup is **for** [MASK].

Both (2a) and (2b) exemplify the 'purpose of item' sense of **for**, for different kinds of items that have different kinds of ordinary purposes. If the pseudoword inferred from the original sentence appropriately generalizes the meaning of **for**, then transplanting it into s' should yield a word like "drinking". However, in most cases the prediction either does not change (here, for example, we still get [MASK] = "reading"), or we get an incoherent prediction for the masked token.

generalization type N (total # of predictions)	**@1** 54	**@5** 270
Vanilla BERT baseline (s' only)	31.5	31.9
MaPP: post hoc average	11.1	14.8
MaPP: aggregate loss (eq. (2))	57.4	53.7

Table 6: Generalization experiment. Comparison of @1 and @5 accuracy, over two generalization types; simple average of the pseudowords versus averaging in the loss function.

We hypothesize that this is because the pseudoword overfit to the original context—that is, it is incapable of representing the desired sense in new contexts (especially if the meaning of the new context crucially affects what should be predicted in the masked slot).

We hypothesize that it is necessary to take multiple contexts into account in order to produce a flexible-context sense-like vector. One possible strategy is to compute a sense-vector as a simple average of the individually-learned pseudoword. We refer to this as **post hoc averaging**.

Another possible strategy is to train each pseudoword on multiple examples with distinct contexts: we replace eq. (1) with an **aggregate loss** that averages over n sentences containing the same focus token with the same sense η:

$$\mathbf{z}^*_\eta = \arg\min_{\mathbf{z}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} ||BERT(\mathbf{z}) - \mathbf{x}^{(i)}_t||^2 \qquad (2)$$

Note that both approaches inject supervision into the process of training the sense vector by specifying which examples correspond to the same sense.

Data. In this experiment we use the **Generalizaton Portion** of the MaPP Dataset, which contains 138 sentences with 5 ambiguous words and 6 distinct senses. For each sense, there are 23 sentences (with 14 sentences used as the training set to compute the averages, and 9 as the test set).

Evaluation. For each sentence we compute two $\mathbf{z}^*$ pseudowords with the two kinds of averaging (post hoc and aggregate loss), and compare their effectiveness at adapting the expected sense for the new context.

In total we evaluate 270 predictions.

Results. Table 6 shows generalization accuracies for the two techniques as well as Vanilla BERT. We see that the aggregate loss technique produces a correct prediction a majority of the time, while the static embedding is less accurate and post hoc averaging of pseudowords performs very poorly. These results support our intuition regarding the possibility to generate a representation for pseudowords that generalizes over different usages of the word. However, an ideal representation of a pseudoword would be one that could serve as a sense-disambiguated embedding of the focus word. This might not be completely achievable, but the representation might be improved in this direction by learning it over a larger more diverse dataset.

6 Discussion

What is a pseudoword? The optimization problem defined in eq. (1) results in a pseudoword $\mathbf{z}^* \in \mathbb{R}^d$. We use the pseudowords as input vectors to the model, although they are not constrained to the 30k vectors in BERT's vocabulary, but may be arbitrary vectors in $\mathbb{R}^d$. We discuss the validity of such an operation in appendix A.4. In practice, we find that many pseudowords behave as sense-disambiguated input vectors. While our goal is not to explore the pseudoword-space for its own sake—pseudowords are a tool to shed light on the geometry and behavior of the BERT-space—our experiments with pseudowords and artificially perturbed pseudowords reveal that the pseudoword-space contains regions that are semantically coherent as inputs to BERT.

Prospects for the MaPP technique. Our dataset is manually curated to control for specific linguistic phenomena. We expect that pseudoword may be less semantically targeted if learned with larger contexts that create more opportunities for confounds. We note also that senses are not necessarily discrete (Erk and McCarthy, 2009), and it would be worthwhile to explore how graded semantic distinctions are represented, as well as underspecified meanings. We are also interested in exploring how BERT represents tokens in sentences that permit multiple plausible interpretations. The MaPP technique can be applied to investigate the properties of other CR models as well, as it requires only that the model be a differentiable function from input token embeddings to contextualized embeddings.

7 Conclusion

We have presented a novel methodology and a dataset for investigating the geometry of the BERT-space, using a traversal technique which allows for a continuation of the input space. We showed that there is substantial regularity in the BERT-space, with regions that correspond to distinct senses. Moreover, we found evidence for "voids" in the space—regions that do not correspond to any intelligible sense. Our technique gives rise to various types of analysis, creating avenues for future work. Immediate directions that we plan to pursue are (a) examining sense representation in longer, naturally occurring sentences, and (b) extending our analysis to a multilingual setting.

Acknowledgments

This work was supported by the Israel Science Foundation (grant no. 929/17). Taelin Karidi was partially supported by a fellowship from the Hebrew University Center for Interdisciplinary Data Science Research. Vivek Srikumar and Yichu Zhou were partly supported by NSF grants #1801446 (SATC) and #1822877 (Cyberlearning) and an award from Verisk Inc. We appreciate feedback from reviewers, Sean Trott, and members of the NERT lab at Georgetown.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *CoRR*, abs/1608.04207.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. *CoRR*, abs/2102.12452.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: a survey. In *IJCAI 2021 (Twenty-Ninth International Joint Conference on Artificial Intelligence)*, volume 5, pages 4330–4338.

Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020. Corpus REDEWIEDERGABE. In *Proceedings of* *the 12th Language Resources and Evaluation Conference*, pages 803–812, Marseille, France. European Language Resources Association.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations (ICLR)*.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda B. Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. *CoRR*, abs/1906.02715.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. 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.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449, Singapore. 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.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, pages 1–52.

William Gale, Kenneth Church, and David Yarowsky. 1992. Work on statistical methods for word sense disambiguation. *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, page 23–25.

Aina Garí Soler and Marianna Apidianaki. 2020. BERT knows Punta Cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualised representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385, Online. Association for Computational Linguistics.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. 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 529–539, Hong Kong, China. Association for Computational Linguistics.

Evan Hernandez and Jacob Andreas. 2021. The low-dimensional linear geometry of contextualized word representations. *ArXiv*, abs/2105.07109.

John Hewitt and Christopher D. Manning. 2019. 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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tianyu Jiang and Ellen Riloff. 2021. Learning prototypical functions for physical artifacts. 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 6941–6951, Online. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual 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)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6792–6812, Online. Association for Computational Linguistics.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, BERT? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290, New York, New York. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196, Melbourne, Australia. Association for Computational Linguistics.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word 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 66–76, Hong Kong, China. Association for Computational Linguistics.

Aviv Slobodkin, Leshem Choshen, and Omri Abend. 2021. Mediators in determining what processing BERT performs first. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 86–93, Online. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations (ICLR)*.

Sean Trott and Benjamin Bergen. 2021. RAW-C: Relatedness of ambiguous words in context (a new lexical resource for English). 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 7077–7087, Online. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Christos Xypolopoulos, Antoine Tixier, and Michalis Vazirgiannis. 2021. Unsupervised word polysemy quantification with multiresolution grids of contextual embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association* *for Computational Linguistics: Main Volume*, pages 3391–3401, Online. Association for Computational Linguistics.

Mengjie Zhao, Philipp Dufter, Yadollah Yaghoobzadeh, and Hinrich Schütze. 2020. Quantifying the contextualization of word representations with semantic class probing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1219–1234, Online. Association for Computational Linguistics.

Yichu Zhou and Vivek Srikumar. 2021. DirectProbe: Studying representations without classifiers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics.

A Appendix

A.1 The MaPP Dataset

We present here the ambiguous words together with their senses, as used in the MaPP Dataset (table 7).

	Basic portion
about	topic, duration
for	duration, recipient, purpose
had	auxiliary past participle, social event, food, medical condition
in	locative, temporal
on	locative, temporal
run	manage, motion
started	electronic device, information source
with	instrument, feeling, accompanier
	Minimal pairs portion
about	topic, duration
for	recipient, purpose
had	auxiliary past participle, social event, food, medical condition
in	locative, temporal
on	locative, temporal
run	manage, motion
started	electronic device, information source
with	instrument, feeling
	Generalization portion
about	topic
for	purpose
had	auxiliary past participle, food
with	feeling, accompanier

Table 7: Ambiguous words and senses for the different portions of the MaPP Dataset.

A.2 Relational Words As a Test Case

We chose to focus our analysis on relational words §4.1. Understanding what is encoded in representations of these words can shed light on some of the open questions regarding the semantic and syntactic knowledge that is encoded in CRs. From prior works on classification of relational words with CRs (e.g., Liu et al., 2019), we know that these differences (in the sense and the form of relational words) are indeed encoded in them. Indeed, in some settings, it is possible to separate groups of them via simple classifiers. However, this is a weak notion of the knowledge that is encoded in the representation. Other work that focused on probing for function words comprehension (Kim et al., 2019) explored whether qualitatively different objectives lead to demonstrably different sentence representations. To understand to what extent (and how) the form of a word versus its sense is encoded in its contextual representation, we have conducted the experiments that we describe in Section 5.

A.3 Our method vs. Other Probing Methods

Our work addresses two basic shortcomings of most probing methods. First, they strongly rely on the probing dataset used to train and evaluate a classifier. Changing the distribution of examples can shift the results of the probing experiment (e.g., Slobodkin et al., 2021). Second, probing methods give an aggregated picture at the population level, and cannot provide insight at the level of individual examples. Our method does not train a classifier, and can provide information at the instance level; it therefore does not rely on aggregation to yield a meaningful conclusion. Rather, it is designed to allow for an interpretable navigation of the BERT space. While our method does allow reporting trends at the population level by aggregation, results can be traced back to the instance level.

A.4 Further Discussion

Transfer learning using BERT. Although BERT is built as a masked language model, it is often being used as a tool for transfer learning; its produced representations are treated as vectors in a continuous space and are being used with great success for various tasks such as POS tagging, NLI, multilingual alignments, prediction of brain activity patterns, and more (Schuster et al., 2019; Gauthier and Levy, 2019; Rogers et al., 2020).

Our method uses pseudowords as input vectors to the model. However, the vectors that are given as an input to BERT are always one of 30k vectors in BERT's base vocabulary, where any other vector is considered "out of vocabulary". BERT was never meant to receive any vector in $\mathbb{R}^d$ as it is defined over a discrete set, yet we are breaking the "discrete-contract" and asking what is the behaviour of the model given the pseudowords and perturbations of them. To the best of our knowledge this is a novel approach to the exploration of BERT. If BERT was treated only as a masked language model then one could claim that there is a certain set of rules that is needed to be followed in order to infer meaningful conclusions from its outputs. In our approach however, we choose adopt a different view – we think about BERT as a sentence encoder; a function from a sequence of strings to sequence of vectors. We claim that in adopting this approach there is no need to constrain the model to a discrete space. Moreover, this "contract" has in fact already been violated, as contextualized representations are often being used for other tasks other than masked language modeling, and therefore the use of pseudowords as inputs is nothing short of a natural continuation of this idea.

More Analysis Results

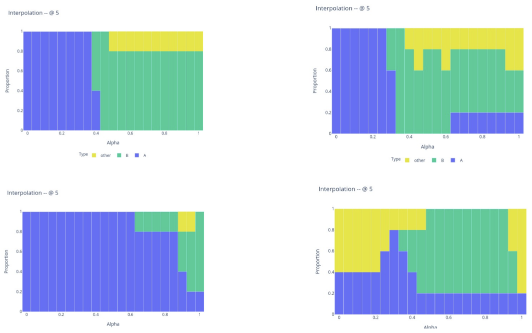

Figure 6: Interpolation results for individual minimal pairs: top-5 sense matches over the range of α values, plotted in the same style as Figure 5. These cases are quite rare and provide examples for departures from the main trend.

	$0 \leq \varepsilon \leq 0.4$		$0.6 \leq \varepsilon \leq 1$		$1.2 \leq \varepsilon \leq 1.8$	
	@1	**@5**	**@1**	**@5**	**@1**	**@5**
N (total # of predictions)	2.82k	4.7k	2.82k	4.7k	3.76k	18.8k
Same sense (z^*)	74.9	63.9	36.3	32.5	14.1	15.7

Table 8: ε-perturbation experiment. Comparison of @1 and @5 accuracy over intervals of ε.

Mask	Vanilla BERT		Query	MaPP							
				$\varepsilon = 0$		$\varepsilon = 0.6$		$\varepsilon = 1.2$		$\varepsilon = 1.8$	
The event is in [MASK].	progress	✗	The event is in London.	London	✓	London	✓	##e	✗	##e	✗
	June	✗		Dublin	✓	Dublin	✓	##a	✗	##ental	✗
	July	✗		Edinburgh	✓	Toronto	✓	##actic	✗	##ated	✗
	April	✗		Paris	✓	Melbourne	✓	##atic	✗	##an	✗
	September	✗		Sydney	✓	Sydney	✓	free	✗	##anche	✗
It lasted for [MASK].	hours	✓	It lasted for seconds.	hours	✓	minutes	✓	##y	✗	forever	✗
	days	✓		minutes	✓	seconds	✓	minutes	✓	hours	✓
	awhile	✓		seconds	✓	hours	✓	long	✗	awhile	✓
	months	✓		awhile	✓	days	✓	##ily	✗	minutes	✓
	weeks	✓		forever	✗	years	✓	seconds	✓	longer	✓
The book is for [MASK].	children	✗	The book is for learning.	reading	✓	children	✗	##o	✗	bilingual	✗
	women	✗		children	✗	reading	✓	##olate	✗	free	✗
	adults	✗		learning	✓	sale	✓	##aged	✗	incomplete	✗
	sale	✓		education	✓	women	✗	##olic	✗	lost	✗
	boys	✗		use	✓	adults	✗	free	✗	anonymous	✗
I started the [MASK].	engine	✓	I started the bus.	engine	✓	car	✓	bird	✗	same	✗
	car	✓		car	✓	truck	✓	phone	✗	bird	✗
	motor	✓		bike	✓	engine	✓	same	✗	hell	✗
	truck	✓		truck	✓	bus	✓	birds	✗	other	✗
	ignition	✓		motor	✓	bike	✓	car	✓	number	✗
I had [MASK].	to	✗	I had slept.	slept	✓	forgotten	✓	asked	✓	asked	✓
	it	✗		forgotten	✓	to	✗	##o	✗	said	✓
	nothing	✗		not	✗	been	✓	nodded	✓	nodded	✓
	him	✗		been	✓	slept	✓	##a	✗	ask	✗
	her	✗		died	✓	##ed	✗	swallowed	✓	smiled	✓
I had a [MASK].	plan	✗	I had a reception.	meeting	✓	lot	✗	lot	✗	little	✗
	point	✗		surprise	✗	smile	✗	little	✗	nod	✗
	headache	✗		party	✓	look	✗	thought	✗	thought	✗
	feeling	✗		call	✗	feeling	✗	feeling	✗	sigh	✗
	choice	✗		date	✓	friend	✗	smile	✗	guess	✗
The clip is about a [MASK].	minute	✗	The clip is about a queen.	woman	✓	woman	✓	shot	✗	video	✓
	year	✗		girl	✓	girl	✓	cartoon	✓	photograph	✓
	second	✗		man	✓	man	✓	song	✓	cartoon	✓
	day	✗		child	✓	mountain	✓	picture	✓	documentary	✓
	week	✗		boy	✓	city	✓	photograph	✓	picture	✓
The clip is about a [MASK].	minute	✓	The clip is about a minute.	minute	✓	minute	✓	documentary	✗	documentary	✗
	year	✓		second	✓	year	✓	ballad	✗	video	✗
	second	✓		third	✗	second	✓	video	✗	photograph	✗
	day	✓		minutes	✗	day	✓	photograph	✗	diary	✗
	week	✓		moment	✓	week	✓	film	✗	ballad	✗
The dinner is on the [MASK].	table	✓	The dinner is on the counter.	table	✓	table	✓	same	✗	same	✗
	rocks	✓		counter	✓	floor	✓	best	✗	following	✗
	way	✗		floor	✓	kitchen	✗	opposite	✗	first	✗
	beach	✓		nightstand	✓	counter	✓	first	✗	winner	✗
	menu	✗		kitchen	✗	menu	✗	truth	✗	result	✗
The dinner is on [MASK].	fire	✗	The dinner is on Monday.	Sunday	✓	free	✗	free	✗	free	✗
	offer	✗		Saturday	✓	open	✗	open	✗	private	✗
	sale	✗		Thursday	✓	served	✗	served	✗	open	✗
	Friday	✓		Tuesday	✓	prepared	✗	closed	✗	served	✗
	hold	✗		Friday	✓	closed	✗	private	✗	delicious	✗

Table 9: Examples top-5 prediction with ε-Perturbubed MaPP versus vanilla BERT. 3 indicated that the prediction has been coded as consistent with Query sense, 7 for wrong prediction. The expectation is that values of ε closer to 0 will be more reflective of the Query sense, while as ε increases will be more incoherent or will not fit the Query sense