What do you mean, BERT?

Anonymous NAACL submission

Abstract

Contextualized word embeddings, i.e. vector representations for words in context, are naturally seen as an extension of previous non-contextual distributional semantic models. In this work, we focus on BERT, a deep neural network that produces contextualized embeddings and has set the state of the art in several semantic tasks, and probe its embedding space for semantic coherence. While showing a tendency towards coherence, BERT does not fully live up to the natural expectations for a semantic vector space. In particular, we find that the position of the sentence in which a word occurs, while having no meaning correlates, leaves a noticeable trace on the word embeddings and disturbs similarity relationships.

1 Introduction

A recent success story of NLP, BERT (Devlin et al., 2018) stands at the crossroad of two key innovations that have brought about significant improvements over previous state of the art results. On the one hand, BERT models are an instance of contextual embeddings (McCann et al., 2017; Peters et al., 2018), which have been shown to be subtle and accurate representations of words within sentences. On the other hand, BERT is a variant of the Transformer architecture (Vaswani et al., 2017) which has set a new state of the art on a wide variety of tasks ranging from machine translation (Ott et al., 2018) to language modeling (Dai et al., 2019). BERT-based models have significantly increased state-of-the-art over the GLUE benchmark for natural language understanding (Wang et al., 2019b) and most of the best scoring models for this benchmark include or elaborate on BERT. Using BERT representations has become in many cases a new standard approach: for instance, all submissions at the recent shared task on gendered pronoun resolution (Webster et al., 2019) were based on BERT. Furthermore, BERT serves both as a strong baseline and as a basis for a fine-tuned state of the art word sense disambiguation pipeline (Wang et al., 2019a).

Analyses aiming to understand the mechanical behavior of Transformers in general, and BERT in particular, have highlighted that they compute word representations through implicitly learned syntactic operations (Raganato and Tiedemann, 2018; Clark et al., 2019; Coenen et al., 2019; Jawahar et al., 2019, a.o.): representations computed through the ‘attention’ mechanisms of Transformers can be seen as weighted sums of intermediary representations from the previous layer, and it has been shown that many attention heads assigned higher weights to syntactically related tokens. Cf. also Linzen et al. (2016); Lakretz et al. (2019) for related studies applied to another neural architecture.

Complementing these previous studies, in this paper we adopt a more theoretically driven lexical semantic perspective. While a clear parallel was established between ‘traditional’ noncontextual embeddings and the theory of distributional semantics (Lenci, 2018; Boleda, 2019, eg.), this link is not automatically extended to contextual embeddings: some authors (Westera and Boleda, 2019) even explicitly consider only “context-invariant” representations as distributional semantics. Hence, we propose to study to what extent BERT, as a contextual embedding architecture, satisfies the properties expected from a natural contextualized extension of distributional semantics models (DSMs).

DSMs are semantic vector spaces: they map each word in the domain of the model to a numeric vector on the basis of distributional criteria. Vector components are inferred from text data. Word properties encoded by DSMs vary from morphological information (Marelli and Baroni, 2015;
Bonami and Paperno, 2018, eg.) to geographic information (Louwerse and Zwaan, 2009), to social stereotypes (Bolukbasi et al., 2016) and to word reference (Herbelot and Vecchi, 2015). A reason why contextualized embeddings have not been equated to distributional semantics may lie in that, as Peters et al. (2018) put it, they are “functions of the entire input sentence”. Contextualized embeddings have introduced some key differences regarding what is to be encoded in a vector representation: whereas traditional DSMs match word types with numeric vectors, contextualized embeddings produce distinct vectors per token. Ideally, the contextualized nature of these embeddings should reflect the semantic nuances that context induces in the meaning of a word—with varying degrees of subtlety, ranging from broad word-sense disambiguation (e.g. ‘bank’ as a river embankment or as a financial institution) to narrower subtypes of word usage (‘bank’ as a corporation or as a physical building) and to more context-specific nuances.

Regardless of how apt contextual embeddings such as BERT are at capturing increasingly finer semantic distinctions, we expect the contextual variation to preserve the basic DSM properties: that the space structure encodes meaning similarity and that variation within the embedding space is semantic in nature. Similar words should be represented with similar vectors, and only semantically pertinent distinctions should affect these representations. We connect our study with previous work in section 2 before detailing the two approaches we followed. First, we verify in section 3 that BERT embeddings form natural clusters when grouped by word types, which on any account should be groups of similar words and thus be assigned similar vectors. Second, we test in sections 4 and 5 that contextualized word vectors do not encode semantically irrelevant features: in particular, leveraging some knowledge from the architectural design of BERT, we address whether there is no systematic difference between BERT representations in odd and even sentences of running text—a property we refer to as cross-sentence coherence. In section 4, we test whether we can observe cross-sentence coherence for single tokens, whereas in section 5 we study to what extent incoherence of representations across sentences affects the similarity structure of the semantic space. We summarize our findings in section 6.

2 Theoretical background & connections

Word embeddings have been said to be ‘all-purpose’ representations, capable of unifying the otherwise heterogeneous domain that is NLP (Turney and Pantel, 2010). To some extent this claim spearheaded the evolution of NLP: focus recently shifted from task-specific architectures with limited applicability to universal architectures requiring little to no adaptation (Radford, 2018; Devlin et al., 2018; Radford et al., 2019; Yang et al., 2019; Liu et al., 2019, a.o.).

Word embeddings are linked to the distributional hypothesis, according to which “you shall know a word from the company it keeps” (Firth, 1957). Accordingly, the meaning of a word can be inferred from the effects it has on its context (Harris, 1954); as this framework equates the meaning of a word to the set its possible usage contexts, it corresponds more to holistic theories of meaning (Quine, 1960, a.o.) than to truth-value accounts (Frege, 1892, a.o.). In early works on word embeddings (Bengio et al., 2003, eg.), a straightforward parallel between word embeddings and distributional semantics could be made: the former are distributed representations of word meaning, the latter a theory stating that word meaning is drawn from distribution. The parallel is much less obvious for contextual embeddings: are constantly changing representations truly an apt description of the meaning of a word?

The analyses that contextual embeddings have been subjected to differ from most analyses of distributional semantics models. Peters et al. (2018) analyzed through an extensive ablation study of ELMo what information is captured by each layer of their architecture. Devlin et al. (2018) discussed what part of their architecture is critical to the performances of BERT, comparing pre-training objectives, number of layers and training duration. Other works (Raganato and Tiedemann, 2018; Hewitt and Manning, 2019; Clark et al., 2019; Voita et al., 2019; Michel et al., 2019) have introduced specific procedures to understand how attention-based architectures function on a mechanical level, generally focusing on either attention matrices or representations across layers. With the exception of Coenen et al. (2019), these analyses leave untouched the question whether BERT contextual embeddings depict a coherent semantic space on their own. Many also involve a learned classifier to ‘extract’ information from the
embeddings, yet this methodology may conflict with the intended purpose of studying the representations themselves (Wieting and Kiela, 2019; Cover, 1965).

The literature on distributional semantics has also put forth and discussed many mathematical properties of embeddings: embeddings are equivalent to count-based matrices (Levy and Goldberg, 2014b), expected to be linearly dependent (Arora et al., 2016), describable as a unitary matrix (Smith et al., 2017) or as a perturbation of an identity matrix (Yin and Shen, 2018). All these properties have however been formalized for non-contextual embeddings: they were formulated using the tools of matrix algebra, under the assumption that embedding matrix rows correspond to word types. Hence they cannot be applied as such to contextualized embeddings. This disconnect in the literature leaves unanswered the question of what consequences there are to framing contextualized embeddings as DSMs.

3 Experiment 1: Word Type Cohesion

The trait of distributional spaces that we focus on in this study is that similar words should lie in similar regions of the semantic space. This should hold all the more so for identical words, which ought to be be maximally similar. By their nature, contextualized embeddings like BERT exhibit variation within vectors corresponding to identical word types. Thus, if BERT is a DSM, we might expect that word types form natural clusters of embeddings that occupy distinctive regions in the embedding space. Here, we assess the coherence of word type clusters by means of their silhouette scores (Rousseeuw, 1987).

3.1 Data & Experimental setup

Throughout our experiments, we used the Gutenberg corpus as provided by the NLTK platform, out of which we removed older texts (the King John’s Bible and Shakespeare). Sentences are enumerated two by two; each pair of sentences is then used as a distinct input source for BERT. As we treat the BERT algorithm as a black box, we retrieve only the embeddings from the last layer, discarding all intermediary representations and attention weights.

To study the basic coherence of BERT’s semantic space, we can consider types as clusters of tokens—i.e. specific instances of contextualized embeddings—and thus leverage the tools of cluster analysis. In particular, silhouette score is generally used to assess whether a specific observation $\vec{v}$ is well assigned to a given cluster $C_i$ drawn from a set of possible clusters $C$. The silhouette score is defined in eq. 1:

$$\text{sep}(\vec{v}, C_i) = \min \{ \text{mean} \, d(\vec{v}, \vec{v}') \, \forall \, C_j \in C - \{C_i\} \}$$

$$\text{coh}(\vec{v}, C_i) = \text{mean} \, d(\vec{v}, \vec{v}') \, \forall \, \vec{v} \in C_i - \{\vec{v}\}$$

$$\text{silh}(\vec{v}, C_i) = \frac{\text{sep}(\vec{v}, C_i) - \text{coh}(\vec{v}, C_i)}{\max\{\text{sep}(\vec{v}, C_i), \text{coh}(\vec{v}, C_i)\}}$$ (1)

We used Euclidean distance for $d$. In our case, observations $\vec{v}$ therefore correspond to tokens, and clusters $C_i$ to types.

Silhouette score consists in computing for each vector observation $\vec{v}$ a cohesion score (viz. the average distance to other observations in the cluster $C_i$) and a separation score (viz. the minimal average distance to other observations if $\vec{v}$ was to be assigned to any other cluster than $C_i$). Optimally, cohesion is to be minimized and separation is to be maximized, and this is reflected in the silhouette score itself: scores are defined between -1 and 1; -1 denotes that the observation $\vec{v}$ should be assigned to another cluster than $C_i$, whereas 1 denotes that the observation $\vec{v}$ is entirely consistent with the cluster $C_i$. Keeping track of silhouette scores for a large number of vectors quickly becomes intractable, hence we use a slightly modified version of the above definition, and compute separation and cohesion using the distance to the average vector for a cluster rather than the average distance to other vectors in a cluster, as suggested by Vendramin et al. (2013). Though results are not entirely equivalent as they ignore the inner structure of clusters, they still present a gross view of the consistency of the vector space under study.

We do note two caveats with our proposed methodology. Firstly, BERT uses subword representations, and thus BERT tokens do not necessarily correspond to words. However we may conjecture that some subwords exhibit coherent meanings, based on whether they tightly correspond to morphemes—e.g. ‘#hs’, ‘#ing’ or ‘#ness’. Secondly, we group word types based on character strings; yet only monosemous words should describe perfectly coherent clusters—whereas we expect some degree of variation for polysemous words and homonyms according to how widely their meanings may vary.
3.2 Results & Discussion

We compared cohesion to separation scores using a paired Student’s t-test, and found a significant effect ($p$-value $< 2 \cdot 2^{-16}$). In short, the effect highlights that cohesion scores are lower than separation scores. The effect size as measured by Cohen’s $d$ (Cohen’s $d = -0.121$) is however rather small, suggesting that cohesion scores are only 12% lower than separation scores. More problematically, we can see in figure 1 that 25.9% of the tokens have a negative silhouette score: one out of four tokens would be better assigned to some other type than the one they belong to. When agglomerating scores by types, we found that 10% of types contained only tokens with negative silhouette score.

Though the overall picture seems satisfying, details show that the standards we expect of DSMs are not always upheld strictly; the median and mean score are respectively at 0.08 and 0.06, indicating a general trend of low scores, even when they are positive. We previously noted that both the use of sub-word representations rather than word representations in BERT and polysemy and homonymy might impact these results. We can quantify the variation in meanings induced by polysemy and homonymy using a dictionary as a sense inventory: the number of distinct entries for a type will give us some insight on how much its meaning varies. We thus used a linear model to predict silhouette scores with log-scaled frequency and log-scaled definition counts, as listed in the Wiktionary, as predictors. We selected tokens for which we found at least one entry in the Wiktionary, out of which we then randomly sampled 10000 observations. Both definition counts and frequency were found to be significant predictors, leading the silhouette score to decrease. This suggests that polysemy degrades the cohesion score of the type cluster, which is compatible with what one would expect from a DSM. We moreover observed that monosemous words yielded higher silhouette scores than polysemous words ($p < 2 \cdot 2^{-16}$, Cohen’s $d = 0.236$), though they still include a substantial number of tokens with negative silhouette scores.

Similarity also includes related words, and not only tokens of the same type. Other studies (Vial et al., 2019; Coenen et al., 2019, eg.) already stressed that BERT embeddings perform well on word-level semantic tasks. To directly assess whether BERT captures this broader notion of similarity, we here used the MEN word similarity dataset (Bruni et al., 2014), which lists pairs of English words with human annotated similarity ratings. We removed pairs containing words for which we had no representation, leaving us with 2290 pairs. We then computed the Spearman correlation between similarity ratings and the cosine of the average BERT embeddings of the two paired word types, and found a correlation of 0.705, showing that cosine similarity of average BERT embeddings encodes semantic similarity.

For comparison, a word2vec DSM (Mikolov et al., 2013a, henceforth w2v) trained on BooksCorpus (Zhu et al., 2015) using the same tokenization as BERT achieved a correlation of 0.669.

4 Experiment 2: Cross-Sentence Coherence

As observed in the previous section, overall the word type coherence in BERT tends to match our basic expectations. In this section, we do further tests, leveraging our knowledge of the design of BERT. We detail the effects of jointly using segment encodings to distinguish between paired input sentences and residual connections, and discuss whether these effects are compatible with a distributional semantics interpretation of these contextualized embeddings.

4.1 Formal approach

We begin by examining the architectural design of BERT. We give some elements relevant to our study here and refer the reader to the original paper by Devlin et al. (2018) for a more complete description. On a formal level, BERT is a deep
neural network composed of superposed layers of computations. Each layer is composed of two sub-layers: the first performing "multi-head attention", the second being a simple feed-forward network. After all sub-layers, residual connections and normalization are applied, thus the intermediary output $\hat{o}_L^i$ after sublayer $L$ can be written as a function of the input $\bar{x}_L^i$, as $\hat{o}_L^i = \text{Norm}(\text{Sub}_L(x_L^i) + x_L^i)$.

BERT is optimized on two training objectives. The first, called masked language model, is a variation on the Cloze test for reading proficiency (Taylor, 1953). The second, called next sentence prediction (NSP), corresponds to predicting whether two sentences are found one next to the other in the original corpus or not. Each example passed as input to BERT is comprised of two sentences, either contiguous sentences from a document, or randomly selected sentences. A special token [SEP] is used to indicate sentence boundaries, and the full sentence is prepended with a second special token [CLS] used to perform the actual prediction for NSP. Each token is transformed into an input vector using an input embedding matrix. To distinguish between tokens from the first and the second sentence, the model adds a learned feature vector $\text{se}\vec{g}_A$ to all tokens from first sentences, and a distinct learned feature vector $\text{se}\vec{g}_B$ to all tokens from second sentences; these feature vectors are called 'segment encodings'. Lastly, as Transformer models do not have an implicit representation of word-order, information regarding the index $i$ of the token in the sentence is added using a positional encoding $p(i)$. Therefore, if the initial training example was "**My dog barks. It is a pooch.**", the actual input would correspond to the following sequence of vectors:

$$[\text{CLS}] + p(\bar{o}) + \text{se}\vec{g}_A, \text{My} + p(\bar{1}) + \text{se}\vec{g}_A,$$

$$\bar{d} + p(\bar{2}) + \text{se}\vec{g}_A, \text{barks} + p(\bar{3}) + \text{se}\vec{g}_A,$$

$$\bar{i} + p(\bar{4}) + \text{se}\vec{g}_A, [\text{SEP}] + p(\bar{5}) + \text{se}\vec{g}_A,$$

$$\bar{\vec{f}} + p(\bar{6}) + \text{se}\vec{g}_B, \bar{i}\vec{s} + p(\bar{7}) + \text{se}\vec{g}_B,$$

$$\bar{\vec{a}} + p(\bar{8}) + \text{se}\vec{g}_B, \text{pooch} + p(\bar{9}) + \text{se}\vec{g}_B,$$

$$\bar{\vec{7}} + p(\bar{10}) + \text{se}\vec{g}_B, [\text{SEP}] + p(\bar{11}) + \text{se}\vec{g}_B$$

Due to the general use of residual connections, marking the sentences using the segment encodings $\text{se}\vec{g}_A$ and $\text{se}\vec{g}_B$ can introduce a systematic offset within sentences. Consider that the first layer uses as input vectors corresponding to word, position, and sentence information: $\bar{w}_i + p(i)$ + $\text{se}\vec{g}_j$; for simplicity, let $\bar{w}_i = \bar{w}_i + p(i)$. The output from the first sub-layer $\bar{o}_1^i$ can be written as:

$$\bar{o}_1^i = \text{Norm}(\text{Sub}_1(\bar{w}_i + \text{se}\vec{g}_j)) + \bar{i}_i + \text{se}\vec{g}_j)$$

$$= \frac{1}{n_1} \text{Sub}_1(\bar{w}_i + \text{se}\vec{g}_j) + \frac{1}{n_1} \bar{i}_i + \frac{1}{n_1} \text{se}\vec{g}_j$$

$$= \bar{o}_1^i + \frac{1}{n_1} \text{se}\vec{g}_j$$

with $n_1 = ||\text{Sub}_1(\bar{w}_i + \text{se}\vec{g}_j) + \bar{i}_i + \text{se}\vec{g}_j||$.

By recurrence, the final output $\bar{o}_L^i$ for a given token $\bar{w}_i + p(i) + \text{se}\vec{g}_j$ can be written as:

$$\bar{o}_L^i = \bar{o}_1^i + \text{se}\vec{g}_j \times \prod_{l=1}^L \frac{1}{n_l}$$

Hence, the segment encoding is partially preserved in the output (i.e. all embeddings within a sentence contain a scaled shift in a specific direction). In principle, this shift may be negligible compared to the remaining part of the output $\bar{o}_L^i$. In figure 2, we illustrate what this systematic shift might entail. Prior to the application of the segment encoding bias, the semantic space is structured by similarity ("pooch" is near "dog") with the bias, we find a different set of characteristics: in our toy example, tokens are linearly separable by sentences.

![Figure 2: Segment encoding bias](image-url)
4.2 Data & Experimental setup

If BERT describes a properly semantic vector space, we should, on average, observe no significant difference in token encoding imputable to the segment the token belongs to. For a given word type \( w \), we may constitute two groups: \( w_{\text{seg}A} \), the set of tokens for this type \( w \) belonging to first sentences in the inputs, and \( w_{\text{seg}B} \), the set of tokens of \( w \) belonging to second sentences. If BERT counterbalances the segment encodings, random differences should cancel out, and therefore the mean of all tokens \( w_{\text{seg}A} \) should be equivalent to the mean of all tokens \( w_{\text{seg}B} \).

We used the same dataset as in section 3. This setting allows us to focus on the effects of the segment encodings. We retrieved the output embeddings of the last BERT layer and grouped them per word type. To assess the consistency of a group of embeddings with respect to a purported reference, we used a mean of squared error (MSE): given a group of embeddings \( E \) and a reference vector \( \vec{\mu} \), we computed how much each vector in \( E \) strayed from the reference \( \vec{\mu} \). It is formally defined as:

\[
\text{MSE}(E, \vec{\mu}) = \frac{1}{\#E} \sum_{i \in E} \sum_d (\vec{v}_d - \vec{\mu}_d)^2 \tag{4}
\]

This MSE can also be understood as the average squared distance to the reference \( \vec{\mu} \). When \( \vec{\mu} = \overline{E} \), i.e. \( \vec{\mu} \) is set to be the average vector in \( E \), the MSE can therefore be likened to a variance in terms of distance. We then used the MSE function to construct pairs of observations: for each word type \( w \), and for each segment encoding \( \text{seg}_i \), we computed two scores:

\[
\text{MSE}(w_{\text{seg}_i}, \overline{w_{\text{seg}_i}}), \quad \text{which gives us an assessment of how coherent the set of embeddings } w_{\text{seg}_i} \text{ is with respect to the mean vector in that set—}
\]

\[
\text{MSE}(w_{\text{seg}_i}, \overline{w_{\text{seg}_j}}), \quad \text{which assesses how coherent the same group of embedding is with respect to the mean vector for the embeddings of the same type, but from the other segment } \text{seg}_j. \text{ If no significant contrast between these two scores can be observed, then BERT counterbalances the segment encodings and is coherent across sentences.}
\]

4.3 Results & Discussion

We compared results using a paired Student’s t-test, which highlighted a significant difference based on which segment types belonged to (p-value \( < 2 \cdot 10^{-16} \)); the effect size (Cohen’s \( d \) = \( -0.527 \)) was found to be stronger than what we computed when assessing whether tokens cluster according to their types (cf. section 3). A visual representation of these results, log-scaled, is shown in figure 3. For all sets \( w_{\text{seg}_i} \), the average embedding from the set itself was systematically a better fit than the average embedding from the paired set \( w_{\text{seg}_j} \). We also noted that a small number of items yielded a disproportionate difference in MSE scores and that frequent word types had smaller differences in MSE scores: roughly speaking, very frequent items—punctuation signs, stopwords, frequent word suffixes—received embeddings that are almost coherent across sentences. Though this effect may be due to segment encodings, we note that BERT uses absolute positional encodings and that the same remark made earlier on the impact of residual connections may apply to these positional encodings as well. Nonetheless, whether we can interpret cross-sentence incoherent embeddings as distributional semantics is a separate question from their exact mechanical cause. We however note that many downstream applications use a single segment encoding per input, de facto avoiding the caveat stressed here.

5 Experiment 3: Sentence-level structure

We previously saw that BERT embeddings do not respect cross-sentence coherence. Perhaps segment encodings have an effect outside of the shift we saw for tokens: they could also impact the relation between any two tokens in a given sentence.

5.1 Data & Experimental setup

Consistent with previous experiments, we used the same dataset (cf. section 3); in this experiment...
also mitigating the impact of the NSP objective (cf. section 4) was crucial. Sentences were thus passed two by two as input to the BERT model. As cosine similarity has been traditionally used to quantify semantic similarity between words (Mikolov et al., 2013b; Levy and Goldberg, 2014a, eg.), we then computed pairwise cosine of the tokens in each sentence. This allows us to reframe our assessment of whether lexical contrasts are coherent across sentences as a comparison of semantic dissimilarity across sentences. More formally, we compute the following set of cosine scores $C_S$ for each sentence $S$:

$$C_S = \{ \cos(\vec{v}, \vec{u}) \mid \vec{v} \neq \vec{u} \land \vec{v}, \vec{u} \in E_S \}$$  \hspace{1cm} (5)$$

with $E_S$ the set of embeddings for the sentence $S$. In this analysis, we compare the union of all sets of cosine scores for first sentences against the union of all sets of cosine scores for second sentences. To avoid asymmetry, we remove the [CLS] token (only present in first sentences), and as with previous experiments we neutralize the effects of the NSP objective by using only consecutive sentences as input.

### 5.2 Results & Discussion

![Figure 4: Wilcoxon tests, 1st vs. 2nd sentences](image)

We compared cosine scores for first and second sentences using a Wilcoxon rank sum test. We observed a significant effect, however small (Cohen’s $d = 0.011$). This may perhaps be due to data idiosyncrasies, and indeed when comparing with a w2V (Mikolov et al., 2013a) trained on BooksCorpus (Zhu et al., 2015) using the same tokenization as BERT, we do observe a significant effect ($p < 0.05$). However the effect size is six times lesser ($d = 0.002$) than what we found for BERT representations; moreover, when varying the sample size (cf. figure 4), $p$-values for BERT representations drop much faster to statistical significance.

A possible reason for the larger discrepancy observed in BERT representations might be that BERT uses absolute positional encodings, i.e. the $k^{th}$ word of the input is encoded with $p(k)$. Therefore, although all first sentences of a given length $l$ will be indexed with the same set of positional encodings $\{p(1), \ldots, p(l)\}$, only second sentences of a given length $l$ preceded by first sentences of a given length $j$ share the exact same set of positional encodings $\{p(j+1), \ldots, p(j+l)\}$. As highlighted previously, the residual connections entailed that the segment encodings were partially preserved in the output embedding: the same argument can be made for positional encodings. In any event, the fact is that we do observe on BERT representations an effect of segment on sentence-level structure. This effect is greater than one can blame on data idiosyncrasies, as verified by the comparison with a traditional DSM such as w2V. If we are to consider BERT as a DSM, we must do so at the cost of cross-sentence coherence.

This may entail that embeddings for tokens drawn from first sentences live in a different semantic space than tokens drawn from second sentences, i.e. that BERT contains two DSMs rather than one. If so, the comparison between two sentence representations from a single input would be meaningless, or at least less coherent than the comparison of two sentence representations drawn from the same sentence position. To test this conjecture, we use two compositional semantics benchmarks: STS (Cer et al., 2017) and SICK-R (Marelli et al., 2014). These datasets are structured as triplets, grouping a pair of sentences with a human-annotated relatedness score. Although the original presentation of BERT (Devlin et al., 2018) did include a downstream application to these datasets, they employed a learned classifier, which complexifies the analysis of the results (Wieting and Kiela, 2019; Cover, 1965). Hence we simply reduce the sequence of tokens within each sentence into a single vector by summing them, a simplistic yet robust semantic composition method. We then compute the Spearman correlation between the cosines of the two sum vectors and the sentence pair’s relatedness score.
| Model                      | STS corr | SICK-R corr |
|---------------------------|----------|-------------|
| Skip-Thought              | 0.25560  | 0.48762     |
| USE                       | 0.66686  | 0.68997     |
| InferSent                 | 0.67646  | 0.70903     |
| **BERT, 2 sent. ipt.**    | 0.35913  | 0.36992     |
| **BERT, 1 sent. ipt.**    | 0.48241  | 0.58695     |
| **w2v**                   | 0.37017  | 0.53356     |

Table 1: Correlation (Spearman $\rho$) of cosine similarity and relatedness ratings on the STS and SICK-R benchmarks.

Results are reported in table 1; baselines include three different sentence encoders and the aforementioned w2v model. As we had suspected, using sum vectors drawn from a single input degrades performances below the w2v baseline. On the other hand, using a single sentence as input seems to produce coherent sentence representations: in that scenario, BERT performs better than w2v and the older sentence encoder Skip-Thought (Kiros et al., 2015), but worse than the modern USE (Cer et al., 2018) and InferSent (Conneau et al., 2017). The comparison with w2v also shows that BERT representations over a coherent input are more likely to include some form of compositional knowledge than traditional DSMs; however it is difficult to decide whether some true form of compositionality is achieved by BERT or whether these performances are entirely a by-product of the positional encodings. In favor of the former, other works have highlighted that Transformer-based architectures perform syntactic operations (Raganato and Tiedemann, 2018; Hewitt and Manning, 2019; Clark et al., 2019; Jawahar et al., 2019; Voita et al., 2019; Michel et al., 2019). In all, these results suggest that the semantic space of token representations from second sentences differ from that of embeddings from first sentences.

6 Conclusions

Our experiments have focused on testing to what extent similar words lie in similar regions of BERT’s latent semantic space. Although we saw that type-level semantics seem to match our general expectations of DSMs, focusing on details leaves us with a much foggier picture.

The main issue stems from BERT’s NSP objective, that requires tokens to be marked according to which sentence they belong. This introduces a distinction between first and second sentence of the input that runs contrary to our expectations in terms of cross-sentence coherence. The validity of such a distinction for lexical semantics may be questioned, yet its effects can be measured. The primary assessment conducted in section 3 shows that token representations did tend to cluster naturally according to their types, yet a finer study detailed in section 4 highlights that tokens from distinct sentences positions tend to have distinct representations. This can be seen as a direct consequence of BERT’s architecture: residual connections, along with the use of specific vectors to encode sentence position, entail that tokens for a given sentence position are ‘shifted’ with respect to tokens for the other position. We also show in section 5 that the use of two sentences as input introduces a variation in lexical contrasts above what can be blamed on corpus specificity.

One way to overcome this violation of cross-sentence coherence would be to consider first and second sentences representations as belonging to distinct distributional semantic spaces. The fact that first sentences were shown to have on average higher pairwise cosines than second sentences can be partially explained by the use of absolute positional encodings in BERT representations. Although positional encodings are required so that the model does not devolve into a bag-of-word system, absolute encodings are not: alternative relative position encodings have been proposed elsewhere in the literature (Shaw et al., 2018; Dai et al., 2019, eg.); replacing the former with the latter may alleviate the gap in lexical contrasts.

Our findings suggest that the formulation of the NSP objective of BERT obfuscates its relation to distributional semantics, by introducing a systematic distinction between first and second sentences which impacts the output embeddings. Similarly, other works (Lample and Conneau, 2019; Yang et al., 2019; Joshi et al., 2019; Liu et al., 2019) stress that the usefulness and pertinence of the NSP task were not obvious. These studies favored an empirical point of view; here, we have shown what sorts of caveats came along with such artificial distinctions from the perspective of a theory of lexical semantics. We hope that future research will extend and refine these findings, and further our understanding of the peculiarities of neural architectures as models of linguistic structure.
References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. Linear algebraic structure of word senses, with applications to polysemy. CoRR, abs/1601.03764.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. J. Mach. Learn. Res., 3:1137–1155.

Gemima Boleda. 2019. Distributional semantics and linguistic theory. CoRR, abs/1905.01896.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. S. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 4349–4357. Curran Associates, Inc.

Olivier Bonami and Denis Paperno. 2018. A characterisation of the inflection-derivation opposition in a distributional vector space. Lingua e Linguaggio.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. J. Artif. Intell. Res., 49:1–47.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomi St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. CoRR, abs/1803.11175.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017, pages 1–14.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look At? An Analysis of BERT’s Attention. arXiv e-prints, page arXiv:1906.04341.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and Measuring the Geometry of BERT. arXiv e-prints, page arXiv:1906.02715.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. CoRR, abs/1705.02364.

Thomas M. Cover. 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans. Electronic Computers, 14(3):326–334.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. CoRR, abs/1901.02860.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

J. R. Firth. 1957. A synopsis of linguistic theory 1930-55. Studies in Linguistic Analysis (special volume of the Philological Society), 1952-59:1–32.

Gottlob Frege. 1892. ber Sinn und Bedeutung. Zeitschrift für Philosophie und philosophische Kritik, 100:25–50.

Zellig Harris. 1954. Distributional structure. Word, 10(23):146–162.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: mapping distributional to model-theoretic semantic spaces. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 22–32. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In NAACL-HLT.

Ganesh Jawahar, Benoît Sagot, and DJamé Seddah. 2019. What does BERT learn about the structure of language? In ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. CoRR, abs/1907.10529.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. CoRR, abs/1506.06726.

Yair Lakretz, German Kruszewski, Theo Desbordes, Diewuke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntactic units in LSTM language models. 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 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. CoRR, abs/1901.07291.

Alessandro Lenci. 2018. Distributional models of word meaning. Annual review of Linguistics, 4:151–171.
Omer Levy and Yoav Goldberg. 2014a. **Linguistic regularities in sparse and explicit word representations.** In Proceedings of the Eighteenth Conference on Computational Natural Language Learning, pages 171–180. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014b. **Neural word embedding as implicit matrix factorization.** In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 2177–2185. Curran Associates, Inc.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *CoRR*, abs/1611.01368.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. **Roberta: A robustly optimized BERT pretraining approach.** *CoRR*, abs/1907.11692.

Max M. Louwerse and Rolf A. Zwaan. 2009. Language encodes geographical information. *Cognitive Science* 33 (2009) 5173.

Marco Marelli and Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological review*, 122 3:485–515.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. **Learned in translation: Contextualized word vectors.** *CoRR*, abs/1708.00107.

Paul Michel, Omer Levy, and Graham Neubig. 2019. **Are Sixteen Heads Really Better than One?** *arXiv e-prints*, page arXiv:1905.10650.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. **Scaling neural machine translation.** *CoRR*, abs/1806.00187.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. **Deep contextualized word representations.** *CoRR*, abs/1802.05365.

William Van Ormann Quine. 1960. *Word And Object.* MIT Press.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Peter Rousseeuw. 1987. *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* *J. Comput. Appl. Math.*, 20(1):53–65.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. **Self-attention with relative position representations.** *CoRR*, abs/1803.02155.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859.

Wilson Taylor. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

Peter D. Turney and Patrick Pantel. 2010. From frequency and to meaning and vector space and models of semantics. In *Journal of Artificial Intelligence Research* 37 (2010) 141-188 *Submitted 10/09; published*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need. CoRR*, abs/1706.03762.

Lucas Vendramin, Pablo A. Jaskowiak, and Ricardo J. G. B. Campello. 2013. On the combination of relative clustering validity criteria. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management, SSDBM*, pages 4:1–4:12, New York, NY, USA. ACM.

Loïc Vial, Benjamin Lécouteux, and Didier Schwab. 2019. Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation. *CoRR*, abs/1905.05677.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *CoRR*, abs/1905.09418.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer
Levy, and Samuel R Bowman. 2019a. Super-glue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Kellie Webster, Marta R. Costa-jussà, Christian Hardmeier, and Will Radford. 2019. Gendered ambiguous pronoun (GAP) shared task at the gender bias in NLP workshop 2019. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Florence, Italy. Association for Computational Linguistics.

Matthijs Westera and Gemma Boleda. 2019. Don’t blame distributional semantics if it can’t do entailment. *CoRR*, abs/1905.07356.

John Wieting and Douwe Kiela. 2019. No training required: Exploring random encoders for sentence classification. *CoRR*, abs/1901.10444.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Zi Yin and Yuanyuan Shen. 2018. On the dimensionality of word embedding. *CoRR*, abs/1812.04224.

Yukun Zhu, Jamie Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

NAACL-HLT 2018 Submission ***. Confidential Review Copy. DO NOT DISTRIBUTE.