# What do you mean, BERT? Assessing BERT as a Distributional Semantics Model

Timothee Mickus, Mathieu Constant, Denis Paperno, Kees van Deemter

# What do you mean, BERT?
# Assessing BERT as a Distributional Semantics Model

**Timothee Mickus**
Université de Lorraine
CNRS, ATILF
tmickus@atilf.fr

**Denis Paperno**
Utrecht University
d.paperno@uu.nl

**Mathieu Constant**
Université de Lorraine
CNRS, ATILF
mconstant@atilf.fr

**Kees van Deemter**
Utrecht University
c.j.vandeemter@uu.nl



## Abstract

Contextualized word embeddings, i.e. vector representations for words in context, are naturally seen as an extension of previous noncontextual 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 study the semantic coherence of its embedding space. 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 suggested 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 arguably be seen as weighted sums of intermediary representations from the previous layer, with many attention heads assigning higher weights to syntactically related tokens (however, contrast with Brunner et al., 2019; Serrano and Smith, 2019).

Complementing these previous studies, in this paper we adopt a more theory-driven lexical semantic perspective. While a clear parallel was established between 'traditional' noncontextual embeddings and the theory of distributional semantics (a.o. Lenci, 2018; Boleda, 2019), 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 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 assume that meaning is derived from use in context. DSMs are nowadays systematically represented using vector spaces (Lenci, 2018). They generally 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. DSMs have also been computed for linguistic items other than words, e.g.,

word senses—based both on meaning inventories (Rothe and Schütze, 2015) and word sense induction techniques (Bartunov et al., 2015)—or meaning exemplars (Reisinger and Mooney, 2010; Erk and Padó, 2010; Reddy et al., 2011). The default approach has however been to produce representations for word types. Word properties encoded by DSMs vary from morphological information (Marelli and Baroni, 2015; Bonami and Paperno, 2018) to geographic information (Louwerse and Zwaan, 2009), to social stereotypes (Bolukbasi et al., 2016) and to referential properties (Herbelot and Vecchi, 2015).

A reason why contextualized embeddings have not been equated to distributional semantics may lie in that they are "functions of the entire input sentence" (Peters et al., 2018). 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. Namely, we expect 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 of 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, e.g.), 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 a word's meaning is drawn from its distribution. In short, word embeddings could be understood as a vector-based implementation of the distributional hypothesis. This parallel is much less obvious for contextual embeddings: are constantly changing representations truly an apt description of the meaning of a word?

More precisely, the literature on distributional semantics has put forth and discussed many mathematical properties of embeddings: embeddings are equivalent to count-based matrices (Levy and Goldberg, 2014b), expected to be linearly dependant (Arora et al., 2016), expressible 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 contextual embeddings. This disconnect in the literature leaves unanswered the question of what consequences there are to framing contextualized embeddings as DSMs.

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 pretraining 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. Recent research has however questioned the pertinence of these attention-based analyses (Serrano and Smith, 2019; Brunner et al., 2019); moreover these works have focused more on the inner workings of the networks than on their adequacy with theories of meaning.

One trait of DSMs that is very often encountered, discussed and exploited in the literature is the fact that the relative positions of embeddings are not random. Early vector space models, by design, required that word with similar meanings lie near one another (Salton et al., 1975); as a consequence, regions of the vectors space describe coherent semantic fields.[1] Despite the importance of this characteristic, the question whether BERT contextual embeddings depict a coherent semantic space on their own has been left mostly untouched by papers focusing on analyzing BERT or Transformers (with some exceptions, e.g. Coenen et al., 2019). Moreover, many analyses of how meaning is represented in attention-based networks or contextual embeddings include "probes" (learned models such as classifiers) as part of their evaluation setup to 'extract' information from the embeddings (Peters et al., 2018; Tang et al., 2018; Coenen et al., 2019; Chang and Chen, 2019, e.g.). Yet this methodology has been criticized as potentially conflicting with the intended purpose of studying the representations themselves (Wieting and Kiela, 2019; Cover, 1965); cf. also Hewitt and Liang (2019) for a discussion. We refrain from using learned probes in favor of a more direct assessment of the coherence of the semantic space.

[1] Vectors encoding contrasts between words are also expected to be coherent (Mikolov et al., 2013b), although this assumption has been subjected to criticism (Linzen, 2016).

## 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 design, contextualized embeddings like BERT exhibit variation within vectors corresponding to identical word types. Thus, if BERT is a DSM, we expect that word types form natural, distinctive clusters 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 (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. We used the `bert-large-uncased` model in all experiments[2]; therefore most of our experiments are done on word-pieces.

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:

$$
\begin{aligned}
sep(\vec{v}, C_i) &= \min\{\underset{\vec{v'} \in C_j}{\text{mean}}\, d(\vec{v}, \vec{v'}) \forall\, C_j \in C - \{C_i\}\} \\
coh(\vec{v}, C_i) &= \underset{\vec{v'} \in C_i - \{\vec{v}\}}{\text{mean}}\, d(\vec{v}, \vec{v'}) \\
silh(\vec{v}, C_i) &= \frac{sep(\vec{v}, C_i) - coh(\vec{v}, C_i)}{\max\{sep(\vec{v}, C_i),\ coh(\vec{v}, C_i)\}} \qquad (1)
\end{aligned}
$$

We used Euclidean distance for $d$. In our case, observations $\vec{v}$ therefore correspond to tokens (that is, *word-piece* tokens), and clusters $C_i$ to types.

[2] Measurements were conducted before the release of the `bert-large-uncased-whole-words` model.

Silhouette scores consist 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, i.e. the minimal 'cost' of assigning $\vec{v}$ 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. '*##s*', '*##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 $< 2 \cdot 2^{-16}$). This 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 aggregating scores by types, we found that 10% of types contained only tokens with negative silhouette score.

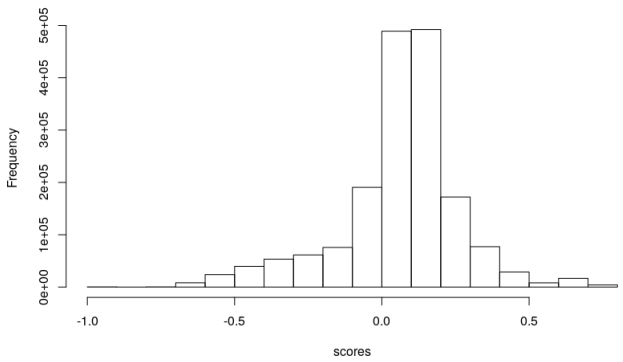



Figure 1: Distribution of token silhouette scores

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 subword representations in BERT as well as polysemy and homonymy might impact these results. The amount of meaning variation induced by polysemy and homonymy can be estimated by using a dictionary as a sense inventory. The number of distinct entries for a type serves as a proxy measure of how much its meaning varies in use. 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, e.g.) already stressed that BERT embeddings perform well on word-level semantic tasks. To directly assess whether BERT captures this broader notion of similarity, we 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*.

### 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 papers by Vaswani et al. (2017) and Devlin et al. (2018), introducing Transformers and BERT, 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. Throughout all layers, after each sub-layer, residual connections and layer normalization are applied, thus the intermediary output $\vec{o_L}$ after sub-layer $L$ can be written as a function of the input $\vec{x_L}$, as $\vec{o_L} = \mathit{LayerNorm}(\mathit{Sub}_L(\vec{x_L}) + \vec{x_L})$.

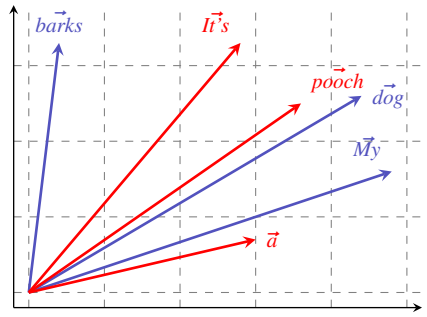
BERT is optimized on two training objectives. The first, called *masked language model*, is a variation on the Cloze test for reading proficiency (Taylor, 1953). The second, called *next sentence prediction* (NSP), corresponds to predicting whether two sentences are found one next to the other in the 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 $\vec{\mathtt{seg}}_A$ to all tokens from first sentences, and a distinct learned feature vector $\vec{\mathtt{seg}}_B$ 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:

$$
\begin{aligned}
&\vec{\mathtt{[CLS]}} + p(\vec{0}) + \vec{\mathtt{seg}}_A, \vec{My} + p(\vec{1}) + \vec{\mathtt{seg}}_A,\\
&\vec{dog} + p(\vec{2}) + \vec{\mathtt{seg}}_A, \vec{barks} + p(\vec{3}) + \vec{\mathtt{seg}}_A,\\
&\vec{.} + p(\vec{4}) + \vec{\mathtt{seg}}_A, \vec{\mathtt{[SEP]}} + p(\vec{5}) + \vec{\mathtt{seg}}_A,\\
&\vec{It} + p(\vec{6}) + \vec{\mathtt{seg}}_B, \vec{is} + p(\vec{7}) + \vec{\mathtt{seg}}_B,\\
&\vec{a} + p(\vec{8}) + \vec{\mathtt{seg}}_B, \vec{pooch} + p(\vec{9}) + \vec{\mathtt{seg}}_B,\\
&\vec{.} + p(\vec{10}) + \vec{\mathtt{seg}}_B, \vec{\mathtt{[SEP]}} + p(\vec{11}) + \vec{\mathtt{seg}}_B
\end{aligned}
$$

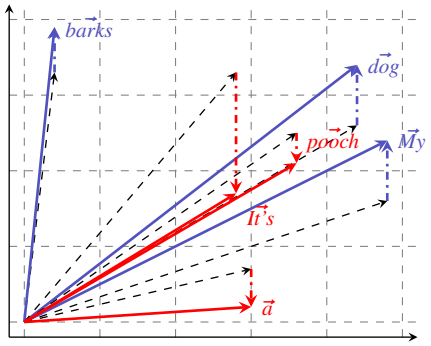
Due to the general use of residual connections, marking the sentences using the segment encodings $\vec{\mathtt{seg}}_A$ and $\vec{\mathtt{seg}}_B$ can introduce a systematic offset within sentences. Consider that the first layer uses as input vectors corresponding to word, position, and sentence information: $\vec{w_i} + \vec{p(i)} + \vec{\mathtt{seg}}_i$; for simplicity, let $\vec{i_i} = \vec{w_i} + \vec{p(i)}$; we also ignore the rest of the input as it does not impact this reformulation. The output from the first sub-layer $\vec{o_i^1}$ can be written:

$$
\begin{aligned}
\vec{o_i^1} &= \mathit{LayerNorm}(\mathit{Sub}_1(\vec{i_i} + \vec{\mathtt{seg}}_i) + \vec{i_i} + \vec{\mathtt{seg}}_i)\\
&= \vec{b_l} + \vec{g^1} \odot \frac{1}{\sigma_i^1}\mathit{Sub}_1(\vec{i_i} + \vec{\mathtt{seg}}_i) + \vec{g^1} \odot \frac{1}{\sigma_i^1}\vec{i_i}\\
&\quad - \vec{g^1} \odot \frac{1}{\sigma_i^1}\mu(\mathit{Sub}_1(\vec{i_i} + \vec{\mathtt{seg}}_i) + \vec{i_i} + \vec{\mathtt{seg}}_i)\\
&\quad + \vec{g^1} \odot \frac{1}{\sigma_i^1}\vec{\mathtt{seg}}_i\\
&= \vec{\tilde{o}_i^1} + \vec{g^1} \odot \frac{1}{\sigma_i^1}\vec{\mathtt{seg}}_i \qquad (2)
\end{aligned}
$$

This equation is obtain by simply injecting the




(a) without segment encoding bias




(b) with segment encoding bias

Figure 2: Segment encoding bias

definition for layer-normalization.[3] Therefore, by recurrence, the final output $\vec{o_i^L}$ for a given token $\vec{w_i} + \vec{p(i)} + \vec{\texttt{seg}_i}$ can be written as:

$$\vec{o_i^L} = \vec{\tilde{o}_i^L} + \left(\bigodot_{l=1}^{L} \vec{g}^l\right) \odot \left(\prod_{l=1}^{L} \frac{1}{\sigma_i^l}\right) \times \vec{\texttt{seg}_i} \quad (3)$$

This rewriting trick shows that segment encodings are partially preserved in the output. All embeddings within a sentence contain a shift in a specific direction, determined only by the initial segment encoding and the learned gain parameters for layer normalization. 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.

[3]Layer normalization after sub-layer $l$ is defined as:

$$\begin{aligned} \textit{LayerNorm}_l(\vec{x}) &= \vec{b_l} + \frac{\vec{g_l} \odot (\vec{x} - \mu(\vec{x}))}{\sigma} \\ &= \vec{b_l} + \vec{g_l} \odot \frac{1}{\sigma}\vec{x} - \vec{g_l} \odot \frac{1}{\sigma}\mu(\vec{x}) \end{aligned}$$

where $\vec{b_l}$ is a bias, $\odot$ denotes element-wise multiplication, $\vec{g_l}$ is a "gain" parameter, $\sigma$ is the standard deviation of components of $\vec{x}$ and $\mu(\vec{x}) = \langle \bar{x}, \ldots, \bar{x} \rangle$ is a vector with all components defined as the mean of components of $\vec{x}$.

### 4.2 Data & Experimental setup

If BERT properly describes a 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: $w_{\texttt{seg}_A}$, the set of tokens for this type $w$ belonging to first sentences in the inputs, and $w_{\texttt{seg}_B}$, 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 $w_{\texttt{seg}_A}$ should be equivalent to the mean of all tokens $w_{\texttt{seg}_B}$.

We used the same dataset as in section 3. This setting (where all paired input sentences are drawn from running text) 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 $\vec{r}$, we computed how much each vector in $E$ strayed from the reference $\vec{r}$. It is formally defined as:

$$\text{MSE}(E, \vec{r}) = \frac{1}{\#E} \sum_{\vec{v} \in E} \sum_{d} (\vec{v}_d - \vec{r}_d)^2 \quad (4)$$

This MSE can also be understood as the average squared distance to the reference $\vec{r}$. When $\vec{r} = \overline{E}$, i.e. $\vec{r}$ is set to be the average vector in $E$, the MSE measures variance of $E$ via Euclidean distance. We then used the MSE function to construct pairs of observations: for each word type $w$, and for each segment encoding $\texttt{seg}_i$, we computed two scores: $\text{MSE}(w_{\texttt{seg}_i}, \overline{w_{\texttt{seg}_i}})$—which gives us an assessment of how coherent the set of embeddings $w_{\texttt{seg}_i}$ is with respect to the mean vector in that set—and $\text{MSE}(w_{\texttt{seg}_i}, \overline{w_{\texttt{seg}_j}})$—which assesses how coherent the same group of embeddings is with respect to the mean vector for the embeddings of the same type, but from the other segment $\texttt{seg}_j$. 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 $< 2 \cdot 2^{-16}$); the effect size (Cohen's $d$ =

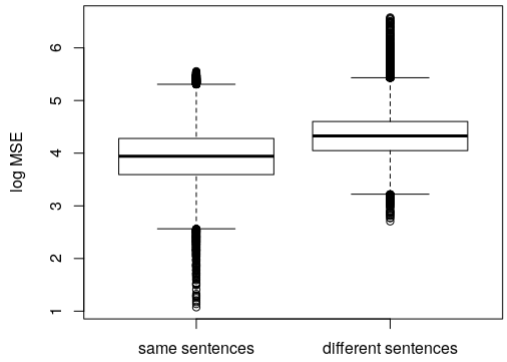



Figure 3: Log-scaled MSE per reference

$-0.527$) 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 $w_{\texttt{seg}_i}$, the average embedding from the set itself was systematically a better fit than the average embedding from the paired set $w_{\texttt{seg}_j}$. 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.

Although the observed positional effect of embeddings' inconsistency might be entirely due to segment encodings, additional factors might be at play. In particular, BERT uses absolute positional encoding vectors to order words within a sequence: the first word $w_1$ is marked with the positional encoding $p(1)$, the second word $w_2$ with $p(1)$, and so on until the last word, $w_n$, marked with $p(n)$. As these positional encodings are added to the word embeddings, the same remark made earlier on the impact of residual connections may apply to these positional encodings as well. Lastly, we also note that many downstream applications use a single segment encoding per input, and thus sidestep the caveat stressed here.

## 5 Experiment 3: Sentence-level structure

We have seen that BERT embeddings do not fully respect cross-sentence coherence; the same type receives somewhat different representations for occurrences in even and odd sentences. However, comparing tokens of the same type in consecutive sentences is not necessarily the main application of BERT and related models. Does the segment-based representational variance affect the structure of the semantic space, instantiated in similarities between tokens of different types? Here we investigate how segment encodings 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 was crucial. Sentences were thus passed two by two as input to the BERT model. As cosine has been traditionally used to quantify semantic similarity between words (Mikolov et al., 2013b; Levy and Goldberg, 2014a, e.g.), 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 $S$:

$$C_S = \{\cos(\vec{v}, \vec{u}) \mid \vec{v} \neq \vec{u} \ \wedge \ \vec{v}, \vec{u} \in E_S\} \quad (5)$$

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

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 smaller ($d = 0.002$) 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.

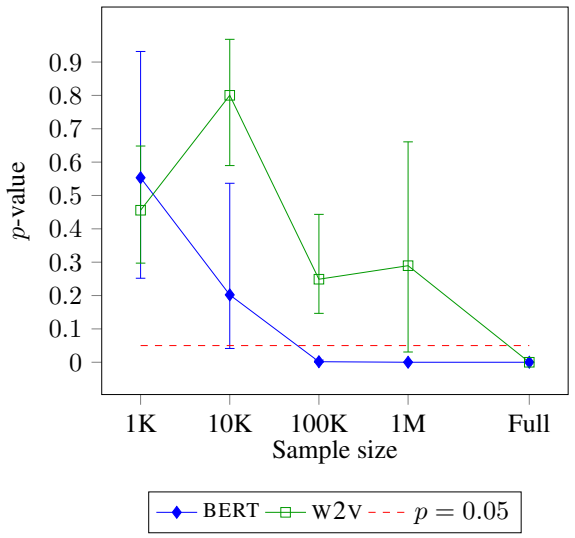



Figure 4: Wilcoxon tests, 1st vs. 2nd sentences

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 ensure 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.

The analysis above suggests 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

| **Model** | **STS cor.** | **SICK-R cor.** |
|---|---|---|
| Skip-Thought | 0.255 60 | 0.487 62 |
| USE | 0.666 86 | 0.689 97 |
| InferSent | 0.676 46 | 0.709 03 |
| BERT, *2 sent. ipt.* | 0.359 13 | 0.369 92 |
| BERT, *1 sent. ipt.* | 0.482 41 | 0.586 95 |
| W2V | 0.370 17 | 0.533 56 |

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

of sentences with a human-annotated relatedness score. The original presentation of BERT (Devlin et al., 2018) did include a downstream application to these datasets, but employed a learned classifier, which obfuscates results (Wieting and Kiela, 2019; Cover, 1965; Hewitt and Liang, 2019). 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. We compare two setups: a "two sentences input" scheme (or *2 sent. ipt.* for short)—where we use the sequences of vectors obtained by passing the two sentences as a single input—and a "one sentence input" scheme (*1 sent. ipt.*)—using two distinct inputs of a single sentence each.

Results are reported in table 1; we also provide comparisons with three different sentence encoders and the aforementioned W2V model. As we had suspected, using sum vectors drawn from a two sentence input scheme single degrades performances below the W2V baseline. On the other hand, a one sentence input scheme 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 research has suggested 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 about DSMs, focusing on details leaves us with a much foggier picture.

The main issue stems from BERT's "next sentence prediction objective", which 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 sentence positions (even vs. odd) tend to have different representations. This can 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. Encodings have a substantial effect on the structure of the semantic subspaces of the two sentences in BERT input. Our experiments demonstrate that assuming sameness of the semantic space across the two input sentences can lead to a significant performance drop in semantic textual similarity.

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: other works have proposed alternative relative position encodings (Shaw et al., 2018; Dai et al., 2019, e.g.); replacing the former with the latter may alleviate the gap in lexical contrasts. Other related questions that we must leave to future works encompass testing on other BERT models such as the whole-words model, or that of Liu et al. (2019) which differs only by its training objectives, as well as other contextual embeddings architectures.

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.

## Acknowledgments

We thank Quentin Gliosca whose remarks have been extremely helpful to this work. We also thank Olivier Bonami as well as three anonymous reviewers for their thoughtful criticism. The work was supported by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program: Idex *Lorraine Université d'Excellence* (reference: ANR-15-IDEX-0004).

## 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.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry P. Vetrov. 2015. Breaking sticks and ambiguities with adaptive skip-gram. *CoRR*, abs/1502.07257.

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

Gemma 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. 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 Langaggio*.

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

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, and Roger Wattenhofer. 2019. On the Validity of Self-Attention as Explanation in Transformer Models. *arXiv e-prints*, page arXiv:1908.04211.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni 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.

Ting-Yun Chang and Yun-Nung Chen. 2019. What does this word mean? explaining contextualized embeddings with natural language definition.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

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. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *ACL*.

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. Über Sinn und Bedeutung. *Zeitschrift fr 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 Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. *arXiv e-prints*, page arXiv:1909.03368.

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 R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

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. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany. Association for Computational Linguistics.

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. *NIPS*.

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. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Belgium, Brussels. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Siva Reddy, Ioannis P. Klapaftis, Diana McCarthy, and Suresh Manandhar. 2011. Dynamic and static prototype vectors for semantic composition. In *IJCNLP*.

Joseph Reisinger and Raymond Mooney. 2010. Multi-prototype vector-space models of word meaning. pages 109–117.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *CoRR*, abs/1507.01127.

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

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

*Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35, Belgium, Brussels. Association for Computational Linguistics.

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, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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 Lecouteux, and Didier Schwab. 2019. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Global Wordnet Conference*, Wroclaw, Poland.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019a. Superglue: 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. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 120–133, Gothenburg, Sweden. Association for Computational Linguistics.

John Wieting and Douwe Kiela. 2019. No training required: Exploring random encoders for sentence classification. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 887–898. Curran Associates, Inc.

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.