What do you mean, BERT?

Assessing BERT as a Distributional Semantics Model

Timothee Mickus
Université de Lorraine
CNRS, ATILF

Denis Paperno
Utrecht University

Mathieu Constant
Université de Lorraine
CNRS, ATILF

Kees van Deemter
Utrecht University

Outline of this talk:

1. Motivation
2. Experiments conducted:
   2.1 Word-type cohesion
   2.2 Cross-sentence coherence
   2.3 Sentence-level structure
3. Recap

Introduction

Introduction

What is BERT

- BERT (Devlin et al., 2018) is a recent, successful model that uses a new architecture ('Transformer' of Vaswani et al., 2017).
  - has been shown to be widely useful: from QA to NLI to NER

Introduction

What is BERT

- BERT (Devlin et al., 2018) is a recent, successful model that uses a new architecture ('Transformer' of Vaswani et al., 2017).
  - has been shown to be widely useful: from QA to NLI to NER

- BERT is an embedding model that produces **contextualized** word embeddings
  - it maps *tokens* to vectors, rather than *types* to vectors.

Introduction

What is BERT

- BERT (Devlin et al., 2018) is a recent, successful model that uses a new architecture (‘Transformer’ of Vaswani et al., 2017).
  - has been shown to be widely useful: from QA to NLI to NER
- BERT is an embedding model that produces **contextualized** word embeddings
  - it maps *tokens* to vectors, rather than *types* to vectors.
- How to assess models such as BERT is an open topic
  - existing methods of analysis of Transformers have their limitations (Serrano and Smith, 2019; Hewitt and Liang, 2019)

Introduction

Embeddings and Distributional Semantics

An important body of work has established that ‘traditional’ non-contextual embeddings are vector space model implementation of distributional semantics (Lenci, 2018).

Introduction

Embeddings and Distributional Semantics

An important body of work has established that 'traditional' non-contextual embeddings are vector space model implementation of distributional semantics (Lenci, 2018).

Vectors can be derived from word distribution, and we can verify whether they capture meaning.

- Non-neural: LSA (Landauer and Dumais, 1997) transforms cooccurrence counts into vectors
- Neural: W2V (Mikolov et al., 2013) trains model to predict words according to their context of occurrence

Introduction

Embeddings and Distributional Semantics

An important body of work has established that ‘traditional’ non-contextual embeddings are vector space model implementation of distributional semantics (Lenci, 2018).

Vectors can be derived from word distribution, and we can verify whether they capture meaning.

- Non-neural: LSA (Landauer and Dumais, 1997) transforms cooccurrence counts into vectors
- Neural: W2V (Mikolov et al., 2013) trains model to predict words according to their context of occurrence
- Testing, eg. on formal analogy (Mikolov, Yih, and Zweig, 2013): 74.0% on ‘semantic’ relations, 60.0% on ‘syntactic’ relations (Garten et al., 2015)

Are BERT embeddings representations of word meaning (in context)?

Are BERT embeddings representations of word meaning (in context)?

- Formal analogy benchmarks supposes we have word-type representations

Are BERT embeddings representations of word meaning (in context)?

- Formal analogy benchmarks supposes we have word-type representations
- Human similarity judgments of word similarity benchmarks supposes we have word-type representations

Introduction

BERT as Distributional Semantics

Are BERT embeddings representations of word meaning (in context)?

- Formal analogy benchmarks supposes we have word-type representations
- Human similarity judgments of word similarity benchmarks supposes we have word-type representations

Focus: **semantically similar words should lie in similar regions of the vector space.**

Introduction

Embeddings used

- dataset: NLTK Gutenberg sample

Introduction

Embeddings used

- dataset: NLTK Gutenberg sample
- model: `BERT large uncased`

Introduction

Embeddings used

- dataset: NLTK Gutenberg sample
- model: `BERT large uncased`
- input format: 2 running sentences of text per input

  no overlap between 1st and 2nd sentences

Word-Type Cohesion

Word-type cohesion

General Intuition

- semantically similar words should have similar embeddings

Word-type cohesion

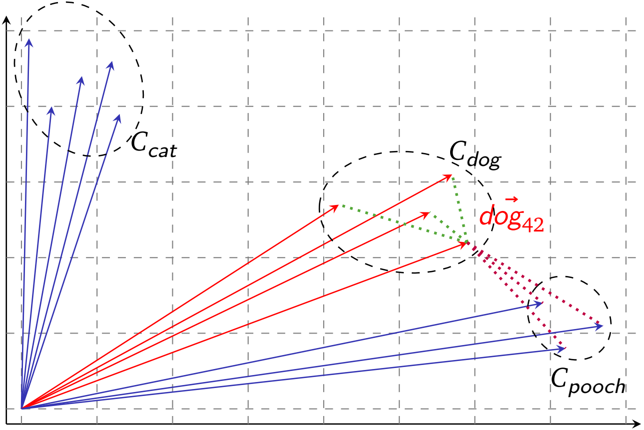
General Intuition

- semantically similar words should have similar embeddings
- so **types should describe coherent clusters of similar token embeddings**

Word-type cohesion

General Intuition

- semantically similar words should have similar embeddings
- so **types should describe coherent clusters of similar token embeddings**
- this can be assessed with silhouette scores (Rousseeuw, 1987)



Word-type cohesion

Experimental Setup

Formally, silhouette scores are defined as:

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

$$\mathit{cohesion}(\vec{v}, C_i) = \underset{\vec{v'} \in C_i - \{\vec{v}\}}{\text{mean}}\ d(\vec{v}, \vec{v'})$$

$$\mathit{silhouette}(\vec{v}, C_i) = \frac{\mathit{separation}(\vec{v}, C_i) - \mathit{cohesion}(\vec{v}, C_i)}{\max\{\mathit{separation}(\vec{v}, C_i),\ \mathit{cohesion}(\vec{v}, C_i)\}}$$

Word-type cohesion

Experimental Setup

Formally, silhouette scores are defined as:

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

$$\mathit{cohesion}(\vec{v}, C_i) = \underset{\vec{v'} \in C_i - \{\vec{v}\}}{\text{mean}}\ d(\vec{v}, \vec{v'})$$

$$\mathit{silhouette}(\vec{v}, C_i) = \frac{\mathit{separation}(\vec{v}, C_i) - \mathit{cohesion}(\vec{v}, C_i)}{\max\{\mathit{separation}(\vec{v}, C_i),\ \mathit{cohesion}(\vec{v}, C_i)\}}$$

Example values:

- If $\mathit{cohesion}(\vec{v}, C_i) = 1$ and $\mathit{separation}(\vec{v}, C_i) = 100$,
  then $\mathit{silhouette}(\vec{v}, C_i) = \frac{100-1}{100} = 0.99$

Word-type cohesion

Experimental Setup

Formally, silhouette scores are defined as:

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

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

$$silhouette(\vec{v}, C_i) = \frac{separation(\vec{v}, C_i) - cohesion(\vec{v}, C_i)}{\max\{separation(\vec{v}, C_i),\ cohesion(\vec{v}, C_i)\}}$$

Example values:

- If $cohesion(\vec{v}, C_i) = 1$ and $separation(\vec{v}, C_i) = 100$,
  then $silhouette(\vec{v}, C_i) = \frac{100-1}{100} = 0.99$
- If $cohesion(\vec{v}, C_i) = separation(\vec{v}, C_i)$,
  then $silhouette(\vec{v}, C_i) = 0$

Word-type cohesion

Experimental Setup

Formally, silhouette scores are defined as:

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

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

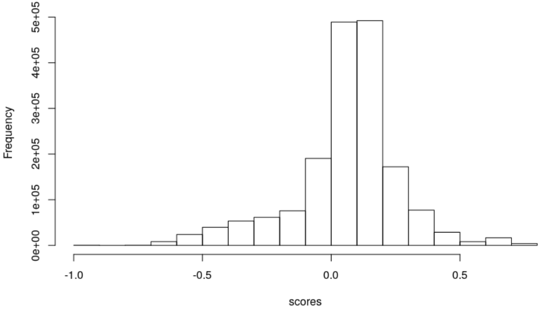
$$silhouette(\vec{v}, C_i) = \frac{separation(\vec{v}, C_i) - cohesion(\vec{v}, C_i)}{\max\{separation(\vec{v}, C_i),\ cohesion(\vec{v}, C_i)\}}$$

Example values:

- If $cohesion(\vec{v}, C_i) = 1$ and $separation(\vec{v}, C_i) = 100$,
  then $silhouette(\vec{v}, C_i) = \frac{100-1}{100} = 0.99$
- If $cohesion(\vec{v}, C_i) = separation(\vec{v}, C_i)$,
  then $silhouette(\vec{v}, C_i) = 0$
- If $cohesion(\vec{v}, C_i) = 100$ and $separation(\vec{v}, C_i) = 1$,
  then $silhouette(\vec{v}, C_i) = \frac{1-100}{100} = -0.99$

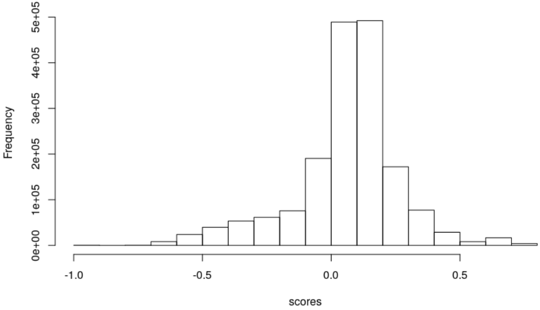
Word-type cohesion

Results



Word-type cohesion

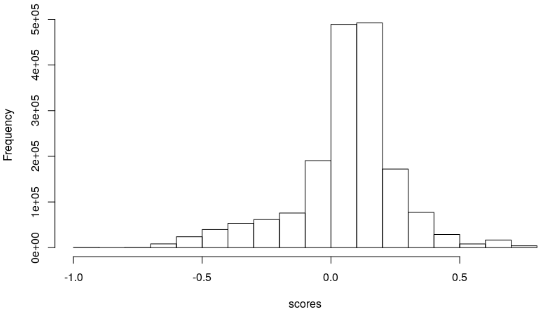
Results



- Overall trend of positive, but low silhouette scores ($25.9\% < 0$)

Word-type cohesion

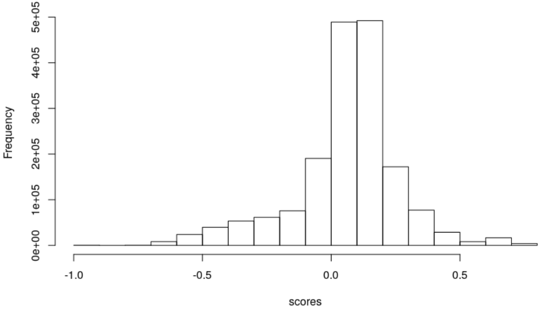
Results



- Overall trend of positive, but low silhouette scores $(25.9\% < 0)$
- Significant difference between cohesion and separation scores $(p < 2 \cdot 2^{-16})$, but rather small effect size $(d = -0.121)$.

Word-type cohesion

Results

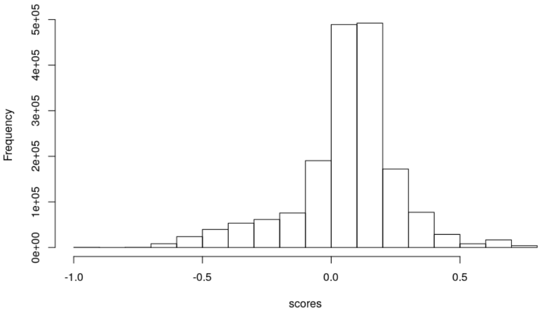


- Overall trend of positive, but low silhouette scores ($25.9\% < 0$)
- Significant difference between cohesion and separation scores ($p < 2 \cdot 2^{-16}$), but rather small effect size ($d = -0.121$).

- Consistently, monosemous word types yield higher silhouette scores than polysemous word types ($d = 0.236$)

Word-type cohesion

Results



- Overall trend of positive, but low silhouette scores ($25.9\% < 0$)
- Significant difference between cohesion and separation scores ($p < 2 \cdot 2^{-16}$), but rather small effect size ($d = -0.121$).

- Consistently, monosemous word types yield higher silhouette scores than polysemous word types ($d = 0.236$)
- Centroids for word types are consistent with human similarity judgment (MEN (Bruni, Tran, and Baroni, 2014): $\rho = 0.705$)

Word-type cohesion
Recap

Overall, word-types are coherently described

Word-type cohesion

Recap

Overall, word-types are coherently described

- The vector space seems consistent with lexical semantics

Word-type cohesion

Recap

Overall, word-types are coherently described

- The vector space seems consistent with lexical semantics
- Some details are not entirely satisfactory

Cross-Sentence Coherence

Cross-Sentence Coherence

General Intuition

- BERT embeddings for tokens are largely semantically cohesive.

Cross-Sentence Coherence

General Intuition

- BERT embeddings for tokens are largely semantically cohesive.
- **Does BERT encode non-semantic information?**

Cross-Sentence Coherence

General Intuition

- BERT embeddings for tokens are largely semantically cohesive.
- **Does BERT encode non-semantic information?**
- For example, BERT architecture suggests that **segment** information could also be encoded.

Cross-Sentence Coherence

Formal Approach: BERT input format

BERT is a Transformer, trained on the "next sentence prediction" objective (viz. '*does the 2^{nd} sentence of the input follow the 1^{st}?*')

Cross-Sentence Coherence

Formal Approach: BERT input format

BERT is a Transformer, trained on the "next sentence prediction" objective (viz. '*does the 2nd sentence of the input follow the 1st?*')

1. tokens are first embedded

Cross-Sentence Coherence

Formal Approach: BERT input format

BERT is a Transformer, trained on the "next sentence prediction" objective (viz. '*does the 2nd sentence of the input follow the 1st?*')

1. tokens are first embedded
2. 'positional encodings' $\vec{p(i)}$ mark the position i of the token

Cross-Sentence Coherence

Formal Approach: BERT input format

BERT is a Transformer, trained on the "next sentence prediction" objective (viz. '*does the 2nd sentence of the input follow the 1st?*')

1. tokens are first embedded
2. 'positional encodings' $\vec{p(i)}$ mark the position i of the token
3. 'Segment encodings' $\vec{\mathtt{seg}}_A$, $\vec{\mathtt{seg}}_B$ mark which sentence tokens belong to

Cross-Sentence Coherence

Formal Approach: BERT input format

BERT is a Transformer, trained on the "next sentence prediction" objective (viz. '*does the 2nd sentence of the input follow the 1st?*')

1. tokens are first embedded
2. 'positional encodings' $\vec{p(i)}$ mark the position i of the token
3. 'Segment encodings' $\vec{\mathtt{seg}}_A$, $\vec{\mathtt{seg}}_B$ mark which sentence tokens belong to
4. 2 special tokens: `[SEP]` for sentence boundaries, `[CLS]` for performing the actual prediction

Cross-Sentence Coherence

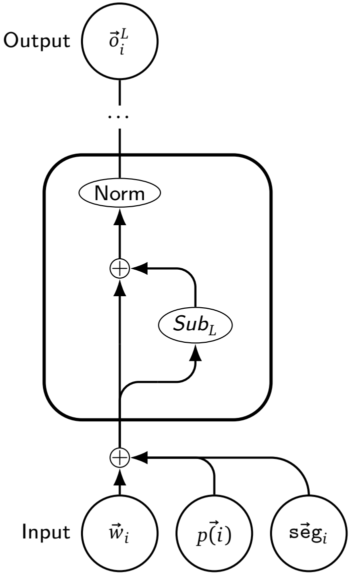
Formal Approach: BERT input format illustrated

Given the example "*My dog barks. It is a pooch.*", the actual input would be:

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

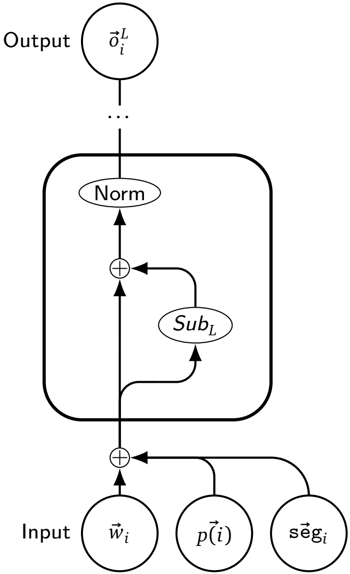
Cross-Sentence Coherence

Formal Approach: BERT architecture



Cross-Sentence Coherence

Formal Approach: BERT architecture

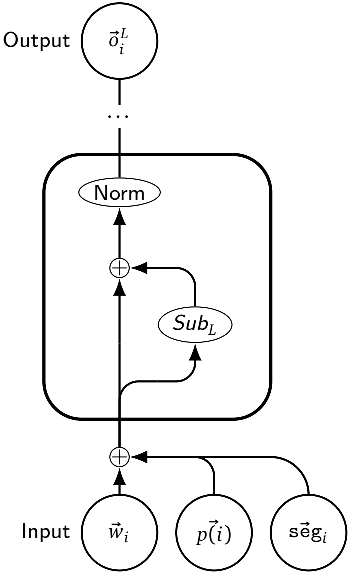


- Let $\vec{i}_i = \vec{w}_i + p(\vec{i})$. The output of the first sublayer of a Transformer (given the input sequence) is:

$$\begin{aligned}
\vec{o}_i^1 &= Norm(Sub_1(\vec{i}_i + \vec{seg}_i) + \vec{i}_i + \vec{seg}_i) \\
&= \vec{b}_l + \vec{g}^1 \odot \frac{1}{\sigma_i^1} Sub_1(\vec{i}_i + \vec{seg}_i) + \vec{g}^1 \odot \frac{1}{\sigma_i^1} \vec{i}_i \\
&\quad - \vec{g}^1 \odot \frac{1}{\sigma_i^1} \mu(Sub_1(\vec{i}_i + \vec{seg}_i) + \vec{i}_i + \vec{seg}_i) \\
&\quad + \vec{g}^1 \odot \frac{1}{\sigma_i^1} \vec{seg}_i \\
&= \tilde{\vec{o}}_i^1 + \vec{g}^1 \odot \frac{1}{\sigma_i^1} \vec{seg}_i
\end{aligned}$$

Cross-Sentence Coherence

Formal Approach: BERT architecture



- Let $\vec{i}_i = \vec{w}_i + p(\vec{i})$. The output of the first sublayer of a Transformer (given the input sequence) is:

$$\begin{aligned}
\vec{o}_i^1 &= Norm(Sub_1(\vec{i}_i + \vec{\text{seg}}_i) + \vec{i}_i + \vec{\text{seg}}_i) \\
&= \vec{b}_l + \vec{g}^1 \odot \frac{1}{\sigma_i^1} Sub_1(\vec{i}_i + \vec{\text{seg}}_i) + \vec{g}^1 \odot \frac{1}{\sigma_i^1} \vec{i}_i \\
&\quad - \vec{g}^1 \odot \frac{1}{\sigma_i^1} \mu(Sub_1(\vec{i}_i + \vec{\text{seg}}_i) + \vec{i}_i + \vec{\text{seg}}_i) \\
&\quad + \vec{g}^1 \odot \frac{1}{\sigma_i^1} \vec{\text{seg}}_i \\
&= \vec{\tilde{o}}_i^1 + \vec{g}^1 \odot \frac{1}{\sigma_i^1} \vec{\text{seg}}_i
\end{aligned}$$

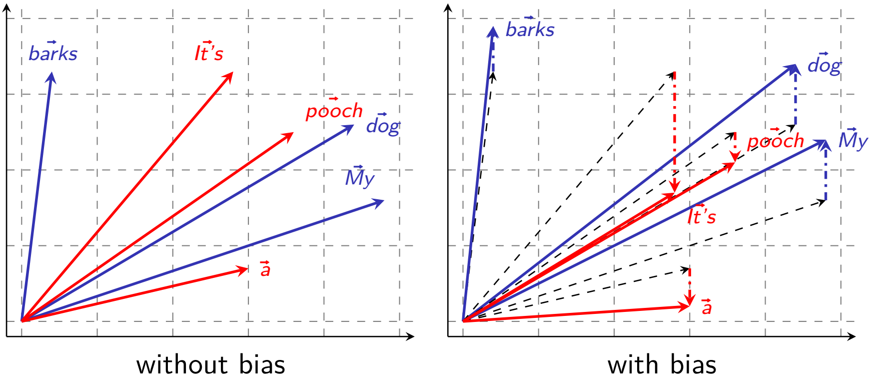
- By recurrence, we get that the final output contains a 'sentence bias':

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

Cross-Sentence Coherence

Visualisation

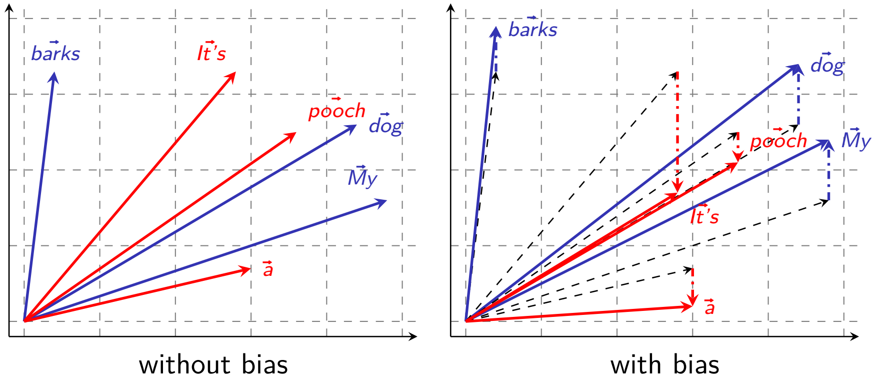
- Each embedding is shifted by a scaled segment encoding



Cross-Sentence Coherence

Visualisation

- Each embedding is shifted by a scaled segment encoding

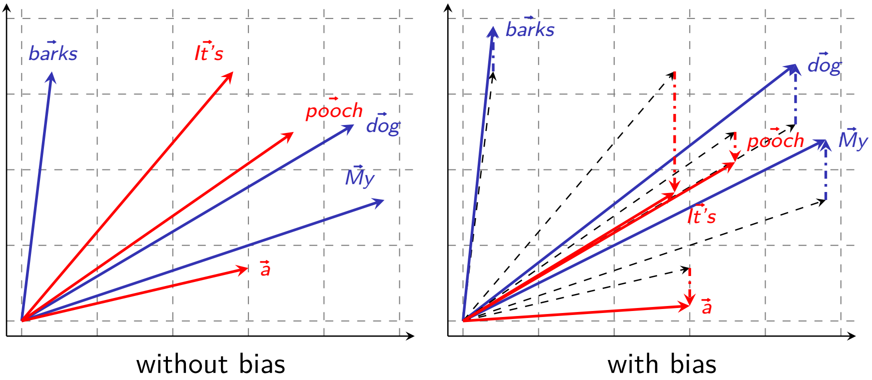


- The bias *may* alter the global characteristics of the embedding space

Cross-Sentence Coherence

Visualisation

- Each embedding is shifted by a scaled segment encoding



- The bias *may* alter the global characteristics of the embedding space
- **Is this bias noticeable?**

Cross-Sentence Coherence

Experimental Setup

- For a given word type w, we constitute two groups:
  1. $w_{\mathtt{seg}_A}$, the set of tokens for this type w belonging to 1st sentences in the inputs
  2. $w_{\mathtt{seg}_B}$, the set of tokens for this type w belonging to 2nd sentences in the inputs

Cross-Sentence Coherence

Experimental Setup

- For a given word type w, we constitute two groups:
  1. $w_{\texttt{seg}_A}$, the set of tokens for this type w belonging to 1st sentences in the inputs
  2. $w_{\texttt{seg}_B}$, the set of tokens for this type w belonging to 2nd sentences in the inputs
- We assess whether the groups are equivalent using mean squared error:

$$\mathrm{MSE}(E, \vec{\mu}) = \frac{1}{\#E} \sum_{\vec{v} \in E} \sum_{d} (\vec{v}_d - \vec{\mu}_d)^2$$

Cross-Sentence Coherence

Experimental Setup

- For a given word type w, we constitute two groups:
  1. $w_{\texttt{seg}_A}$, the set of tokens for this type w belonging to 1st sentences in the inputs
  2. $w_{\texttt{seg}_B}$, the set of tokens for this type w belonging to 2nd sentences in the inputs
- We assess whether the groups are equivalent using mean squared error:

$$\mathrm{MSE}(E, \vec{\mu}) = \frac{1}{\#E} \sum_{\vec{v} \in E} \sum_{d} (\vec{v}_d - \vec{\mu}_d)^2$$

- MSE measures how much observations $\vec{v} \in E$ deviate from the mean value $\vec{\mu}$

Cross-Sentence Coherence

Experimental Setup

- Null hypothesis: segment shift is negligible.

Cross-Sentence Coherence

Experimental Setup

- Null hypothesis: segment shift is negligible.
- If so, $w_{\texttt{seg}_A}$ and $w_{\texttt{seg}_B}$ contain equivalent embeddings,

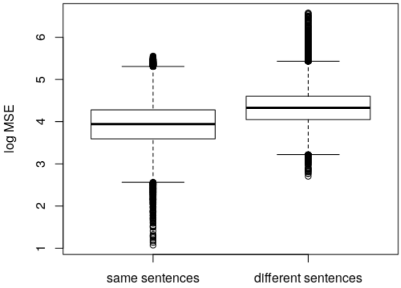
Cross-Sentence Coherence

Experimental Setup

- Null hypothesis: segment shift is negligible.
- If so, $w_{\mathtt{seg}_A}$ and $w_{\mathtt{seg}_B}$ contain equivalent embeddings,
- If so, $\forall \vec{w} \in w_{\mathtt{seg}_i}, \quad \mathrm{MSE}(\{\vec{w}\}, \overline{w_{\mathtt{seg}_i}}) \approx \mathrm{MSE}(\{\vec{w}\}, \overline{w_{\mathtt{seg}_j}})$.

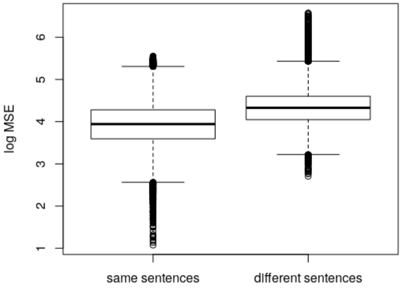
Cross-Sentence Coherence

Results



Cross-Sentence Coherence

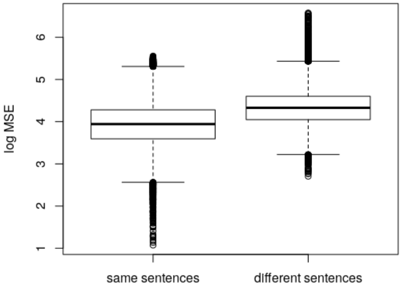
Results



- MSE scores systematically favor the mean of the token's own segment

Cross-Sentence Coherence

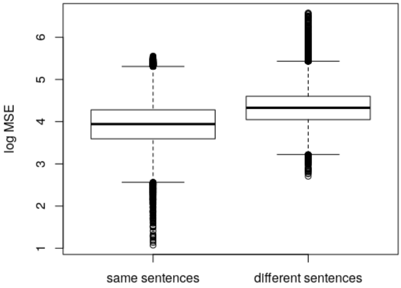
Results



- MSE scores systematically favor the mean of the token's own segment
- Significant difference between segments ($p < 2 \cdot 2^{-16}$), non-negligible effect size ($d = -0.527$).

Cross-Sentence Coherence

Results

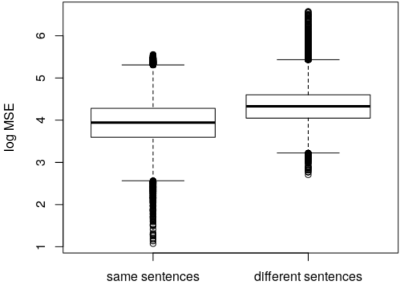


- MSE scores systematically favor the mean of the token's own segment
- Significant difference between segments ($p < 2 \cdot 2^{-16}$), non-negligible effect size ($d = -0.527$).

- Very frequent word types generally have almost segment-insensitive representations

Cross-Sentence Coherence

Results

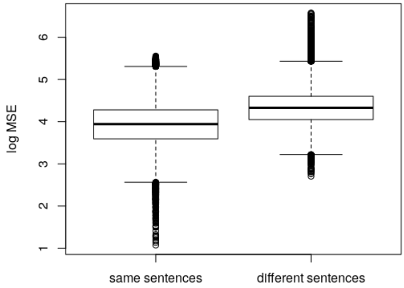


- MSE scores systematically favor the mean of the token's own segment
- Significant difference between segments ($p < 2 \cdot 2^{-16}$), non-negligible effect size ($d = -0.527$).

- Very frequent word types generally have almost segment-insensitive representations
- Rare word types yield high MSE scores

Cross-Sentence Coherence

Results



- MSE scores systematically favor the mean of the token's own segment
- Significant difference between segments ($p < 2 \cdot 2^{-16}$), non-negligible effect size ($d = -0.527$).

- Very frequent word types generally have almost segment-insensitive representations
- Rare word types yield high MSE scores

Frequency plays a role, though it's not the only factor

Cross-Sentence Coherence

Recap

- BERT's architecture encodes a segment based shift, or 'sentence bias'

Cross-Sentence Coherence

Recap

- BERT's architecture encodes a segment based shift, or 'sentence bias'

- **This bias is statistically significant and noticeable**

Cross-Sentence Coherence

Recap

- BERT's architecture encodes a segment based shift, or 'sentence bias'
- **This bias is statistically significant and noticeable**
- BERT embeddings encode at least segment information in addition to semantic information

Sentence-Level Structure

Does the segment bias affect the similarity structure of the embedding space?

Sentence-Level Structure

General Intuition

Does the segment bias affect the similarity structure of the embedding space?

- Semantic similarity scores in 1st and 2nd sentences should be equivalent

Sentence-Level Structure

General Intuition

Does the segment bias affect the similarity structure of the embedding space?

- Semantic similarity scores in 1st and 2nd sentences should be equivalent

  Cosines for any two words in the sentence "I walk my dog." should be more or less equal to cosines for any two words in the sentence "It is a pooch."

Does the segment bias affect the similarity structure of the embedding space?

- Semantic similarity scores in 1st and 2nd sentences should be equivalent

  Cosines for any two words in the sentence "I walk my dog." should be more or less equal to cosines for any two words in the sentence "It is a pooch."

- We shouldn't be able to tell whether some score s corresponds to embeddings drawn from a 1st or a 2nd sentence

Sentence-Level Structure

Experimental Setup

Semantic similarity in DSMs is often assessed using cosines.

Sentence-Level Structure

Experimental Setup

Semantic similarity in DSMs is often assessed using cosines.

- We compute pairwise cosine for each sentence S:

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

Sentence-Level Structure

Experimental Setup

Semantic similarity in DSMs is often assessed using cosines.

- We compute pairwise cosine for each sentence S:

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

- We bin cosines according to their position in the input:

$$C_{\mathtt{seg}_A} = \bigcup_{S \in \mathit{1}^{st} \ \mathit{sent.}} C_S$$

$$C_{\mathtt{seg}_B} = \bigcup_{S \in \mathit{2}^{nd} \ \mathit{sent.}} C_S$$

Sentence-Level Structure

Experimental Setup

Semantic similarity in DSMs is often assessed using cosines.

- We compute pairwise cosine for each sentence S:

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

- We bin cosines according to their position in the input:

$$C_{\mathtt{seg}_A} = \bigcup_{S \in \mathit{1^{st}\ sent.}} C_S$$

$$C_{\mathtt{seg}_B} = \bigcup_{S \in \mathit{2^{nd}\ sent.}} C_S$$

- We compare $C_{\mathtt{seg}_A}$ and $C_{\mathtt{seg}_B}$ with a Wilcoxon rank sum test

Sentence-Level Structure

Experimental Setup

Semantic similarity in DSMs is often assessed using cosines.

- We compute pairwise cosine for each sentence S:

$$C_S = \{\cos(\vec{v}, \vec{u}) \mid \vec{v} \neq \vec{u} \;\wedge\; \vec{v}, \vec{u} \in S\}$$

- We bin cosines according to their position in the input:

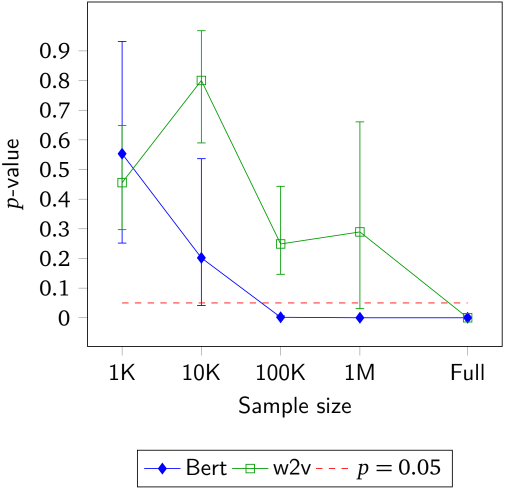
$$C_{\texttt{seg}_A} = \bigcup_{S \in \textit{1}^{\textit{st}} \textit{ sent.}} C_S$$

$$C_{\texttt{seg}_B} = \bigcup_{S \in \textit{2}^{\textit{nd}} \textit{ sent.}} C_S$$

- We compare $C_{\texttt{seg}_A}$ and $C_{\texttt{seg}_B}$ with a Wilcoxon rank sum test
- We also report a W2V baseline

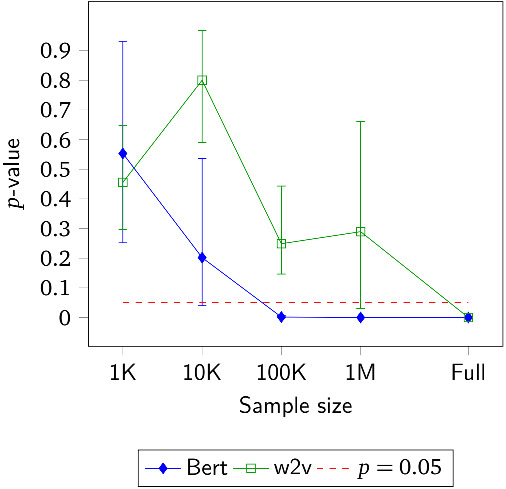
Sentence-Level Structure

Results



Sentence-Level Structure

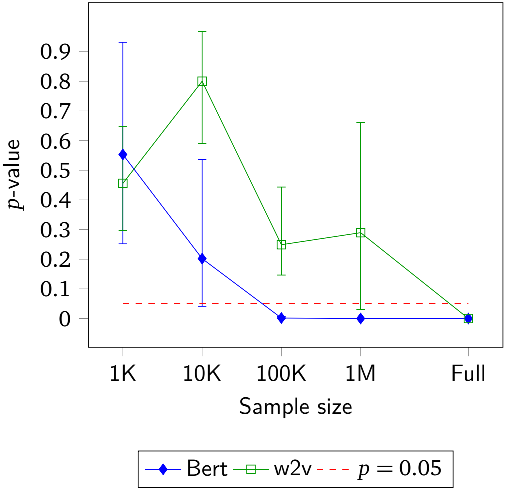
Results



- Significant effect in BERT, small effect size ($d = 0.011$); also found for W2V, even smaller size ($d = 0.002$).

Sentence-Level Structure

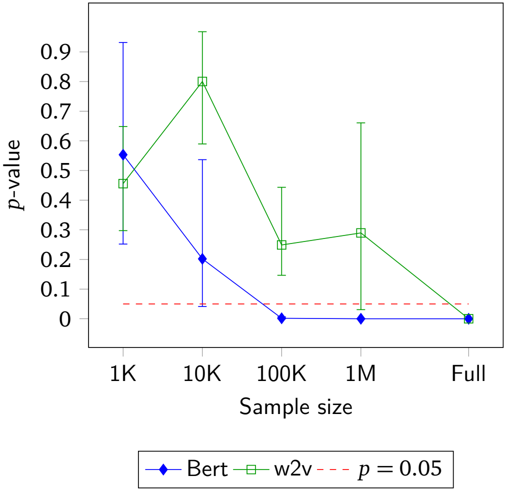
Results



- Significant effect in BERT, small effect size ($d = 0.011$); also found for W2V, even smaller size ($d = 0.002$).
- p-values for BERT random samples are more often significant

Sentence-Level Structure

Results



- Significant effect in BERT, small effect size ($d = 0.011$); also found for W2V, even smaller size ($d = 0.002$).
- p-values for BERT random samples are more often significant
- The effect can't be blamed only on the dataset.

There is a difference of semantic similarity between 1st and 2nd sentences.

There is a difference of semantic similarity between 1st and 2nd sentences.

- On average, similarities of tokens within 1st sentences are greater than similarities of tokens within 2nd sentences.

Sentence-Level Structure

Recap

There is a difference of semantic similarity between 1st and 2nd sentences.

- On average, similarities of tokens within 1st sentences are greater than similarities of tokens within 2nd sentences.

- This difference cannot be semantic because odd vs. even numbers of segments are essentially arbitrary

Conclusions

Conclusions

- We presented tools for the interpretation of BERT, and showed that

Conclusions

- We presented tools for the interpretation of BERT, and showed that
  1. **BERT is generally coherent with respect to lexical semantics**

Conclusions

- We presented tools for the interpretation of BERT, and showed that
  1. **BERT is generally coherent with respect to lexical semantics**
  2. **BERT also encode non-semantic factors**, viz. the difference between 1^{st} and 2^{nd} sentences

Conclusions

- We presented tools for the interpretation of BERT, and showed that
  1. **BERT is generally coherent with respect to lexical semantics**
  2. **BERT also encode non-semantic factors**, viz. the difference between 1st and 2nd sentences
  3. **These factors impact similarity measures**

Conclusions

- We presented tools for the interpretation of BERT, and showed that
  1. **BERT is generally coherent with respect to lexical semantics**
  2. **BERT also encode non-semantic factors**, viz. the difference between 1^{st} and 2^{nd} sentences
  3. **These factors impact similarity measures**

- Linguistically coherent formalizations can prevent that kind of behavior

Conclusions

- We presented tools for the interpretation of BERT, and showed that
  1. **BERT is generally coherent with respect to lexical semantics**
  2. **BERT also encode non-semantic factors**, viz. the difference between 1^{st} and 2^{nd} sentences
  3. **These factors impact similarity measures**

- Linguistically coherent formalizations can prevent that kind of behavior

- In future work:

Conclusions

- We presented tools for the interpretation of BERT, and showed that
  1. **BERT is generally coherent with respect to lexical semantics**
  2. **BERT also encode non-semantic factors**, viz. the difference between 1^{st} and 2^{nd} sentences
  3. **These factors impact similarity measures**

- Linguistically coherent formalizations can prevent that kind of behavior

- In future work:
  1. test other contextualized embeddings (ELMo, Roberta...)

Conclusions

- We presented tools for the interpretation of BERT, and showed that
  1. **BERT is generally coherent with respect to lexical semantics**
  2. **BERT also encode non-semantic factors**, viz. the difference between 1^{st} and 2^{nd} sentences
  3. **These factors impact similarity measures**

- Linguistically coherent formalizations can prevent that kind of behavior

- In future work:
  1. test other contextualized embeddings (ELMo, Roberta...)
  2. focus on other potential non-semantic factors (eg. positional encodings)

References I

Bruni, Elia, Nam-Khanh Tran, and Marco Baroni (2014). "Multimodal Distributional Semantics". In: *J. Artif. Intell. Res.* 49, pp. 1–47.

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805. url: http://arxiv.org/abs/1810.04805.

Garten, Justin et al. (2015). "Combining Distributed Vector Representations for Words". In: *VS@HLT-NAACL*.

Hewitt, John and Percy Liang (2019). "Designing and Interpreting Probes with Control Tasks". In: *arXiv e-prints*, arXiv:1909.03368, arXiv:1909.03368. arXiv: 1909.03368 [cs.CL].

Landauer, Thomas K and Susan T. Dumais (1997). "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge". In: *Psychological Review 1997, Vol. 104*.

Lenci, Alessandro (2018). "Distributional models of word meaning". In: *Annual review of Linguistics* 4, pp. 151–171.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic Regularities in Continuous Space Word Representations.". In: *HLT-NAACL*, pp. 746–751.

References II

Mikolov, Tomas et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781. arXiv: 1301.3781. url: http://arxiv.org/abs/1301.3781.

Rousseeuw, Peter (Nov. 1987). "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". In: *J. Comput. Appl. Math.* 20.1, pp. 53–65. issn: 0377-0427. doi: 10.1016/0377-0427(87)90125-7. url: http://dx.doi.org/10.1016/0377-0427(87)90125-7.

Serrano, Sofia and Noah A. Smith (July 2019). "Is Attention Interpretable?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2931–2951. doi: 10.18653/v1/P19-1282. url: https://www.aclweb.org/anthology/P19-1282.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 5998–6008. url: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.