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

Taelin Karidi, Yichu Zhou, Nathan Schneider, Omri Abend, Vivek Srikumar

Hebrew University of Jerusalem # University of Utah # Georgetown University

EMNLP 2021
Motivating Question

What knowledge is encoded in LMs such as BERT?
Motivating Question

What knowledge is encoded in LMs such as BERT?

- **Downstream model** (e.g. probing classifier)
  - [Liu et al., 2019; Conneau et al., 2018; Belinkov et al., 2017; Adi et al., 2016, inter alia]

- **Predictions of the LM itself**
  - [Petroni et al., 2019]

- **Geometric methods** (e.g. clustering the embeddings)
  - [Coenen et al., 2019; Ethayarajh 2019; Gessler & Schneider 2021]
Motivating Question

What knowledge is encoded in LMs such as BERT?

What does BERT know about **lexical semantics**?
Motivating Question

What knowledge is encoded in LMs such as BERT?

What does BERT know about **lexical semantics**?

highly **ambiguous words** & their **senses**
Motivating Question

What knowledge is encoded in LMs such as BERT?

What does BERT know about **lexical semantics**?

highly **ambiguous words** & their **senses**

- prepositions
- verbs
Motivating Example

The event is in _______
Motivating Example

The event is in ________

ambiguous word What is the sense?
Motivating Example

The event is in ______

The event is in London

The event is in October
Motivating Example

The event is in ________

The event is in London
- locative sense
- disambiguating token

The event is in October
- temporal sense
- disambiguating token

ambiguous word
What is the sense?
Motivating Example

The event is in _______

The event is in London

locative sense

disambiguating token

The event is in October

temporal sense

disambiguating token

How is the information about the sense encoded in the contextualised representation of “in”? 
Main Hypothesis

There are regular “nicely defined” regions in the BERT-space around words that correspond to distinct senses.
Geometric View

**Naive approach:** Look at neighborhoods of points in the BERT-space.

The event is *in London.*
Geometric View

Naive approach: Look at neighborhoods of points in the BERT-space.

Problem: How can we interpret an arbitrary point in the BERT-space?

The event is in London.
Masked Pseudoword Probing (MaPP)

Novel technique to investigate the geometry of the BERT-space in a controlled manner around individual instances.
Masked Pseudoword Probing (MaPP)

Novel technique to investigate the geometry of the BERT-space in a controlled manner around individual instances.

How do different regions in the contextualised space correspond to word senses?
Masked Pseudoword Probing (MaPP)

BERT space (contextualized embedding)

The event is in London

input space (static embeddings)
Masked Pseudoword Probing (MaPP)

We can also use BERT for masked prediction

BERT space
(contextualized embedding)

input space
(static embeddings)

[MASK] =
{ progress
  June
  July
  April
  September
}

The event is in [MASK]

Z₁ Z₂ Z₃ Z₄
We learn a **pseudoword** $z^*$ in place of $z^4_4$ which is customized to reconstruct $x^4_4$.

$$z^* = \arg \min_{z \in \mathbb{R}^d} \| BERT(z) - x_t \|^2$$

(Here $t=4$, $d=768$)
Masked Pseudoword Probing (MaPP)

Masked prediction using a pseudoword

BERT space (contextualized embedding)


input space (static embeddings)

\[ z_1 \quad z_2 \quad z_3 \quad z^* \]

The event is in [MASK]
Masked Pseudoword Probing (MaPP)

BERT space (contextualized embedding)

\[ \text{[MASK]} = \{ \text{London, Dublin, Edinburgh, Paris, Sydney} \} \]

input space (static embeddings)

The event is in [MASK]

\[ z_1 \quad z_2 \quad z_3 \quad z^* \]
Masked Pseudoword Probing (MaPP)

BERT space (contextualized embedding)

The event is in [MASK]

input space (static embeddings)

$z_1$, $z_2$, $z_3$, $z^*$

original static embedding

{London, Dublin, Edinburgh, Paris, Sydney} ✓

{progress, June, July, April, September} ✗
Masked Pseudoword Probing (MaPP): Summary

1. Run BERT for a sentence.

\[ BERT(\text{The event is in October.}) \]
static inputs: \( z_{\text{The}} \ z_{\text{event}} \ z_{\text{is}} \ z_{\text{in}} \ldots \)
contextualized outputs: \( x_{\text{The}} \ x_{\text{event}} \ x_{\text{is}} \ x_{\text{in}} \ldots \)

2. Learn pseudoword \( z_{\text{in}}^* \) in place of \( z_{\text{in}} \) that is customized to reconstruct \( x_{\text{in}} \).

\[ z_{\text{in}}^* = \arg \min_{z \in \mathbb{R}^d} \| BERT(z) - x_{\text{in}} \|^2 \]

3. Run masked prediction with the modified input vector.

\[ BERT(\text{The event is in [MASK].}) \]

4. Examine top predictions for sense match.

October ✔  July ✔  winter ✔  London ✗  #ic ✗

How do we interpret the pseudowords?
The MaPP Data Set

- We manually compiled a dataset for our experiments.
- Each sentence contains an ambiguous word that is fully disambiguated by a specific slot in the sentence.

<table>
<thead>
<tr>
<th>Focus Word</th>
<th>Sentence</th>
<th>Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>in</td>
<td>The event is <strong>in October</strong>.</td>
<td><strong>temporal</strong></td>
</tr>
<tr>
<td>for</td>
<td>The book is <strong>for Lisa</strong>.</td>
<td><strong>person</strong></td>
</tr>
<tr>
<td>with</td>
<td>I ate salad <strong>with enjoyment</strong>.</td>
<td><strong>feeling</strong></td>
</tr>
<tr>
<td>about</td>
<td>The clip is <strong>about a horse</strong>.</td>
<td><strong>topic</strong></td>
</tr>
<tr>
<td>started</td>
<td>I <strong>started</strong> the <strong>car</strong>.</td>
<td><strong>device</strong></td>
</tr>
<tr>
<td>had</td>
<td>I <strong>had</strong> a <strong>party</strong>.</td>
<td><strong>social event</strong></td>
</tr>
<tr>
<td>had</td>
<td>I <strong>had slept</strong>.</td>
<td><strong>auxiliary/past participle</strong></td>
</tr>
</tbody>
</table>
Research Questions & Experiments

**Experiment 1: Specialization**

**Question:** Does a pseudoword decode to a specific sense of the focus token?
Research Questions & Experiments

**Experiment 1: Specialization**

**Question:** Does a pseudoword decode to a specific sense of the focus token?

<table>
<thead>
<tr>
<th>Query</th>
<th>Top 5 predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>The dinner is on Monday</td>
<td>z fire x offer x sale x Friday ✓ hold x</td>
</tr>
<tr>
<td></td>
<td>z* Sunday ✓ Saturday ✓ Thursday ✓ Tuesday ✓ Friday ✓</td>
</tr>
<tr>
<td>The clip is about a queen</td>
<td>z* minute x year x second x day x week x</td>
</tr>
<tr>
<td></td>
<td>z* woman ✓ girl ✓ man ✓ child ✓ boy ✓</td>
</tr>
</tbody>
</table>
Specialization Experiment: Results

Accuracy at producing a completion consistent with the sense from the original context

- Vinalla BERT
- MaPP

Bar chart showing accuracy for different categories and different numbers of completions.
Research Questions & Experiments: Interpolation

Experiment 2: Interpolation

**Question:** What does a boundary between two distinct senses look like?

The event is *in* London

\[ z_\alpha = (1 - \alpha)z_1^* + \alpha z_2^* \]

The event is *in* October
Examples for the interpolation results:

<table>
<thead>
<tr>
<th>Mask</th>
<th>Vanilla BERT</th>
<th>Query 1</th>
<th>Interpolated MaPP</th>
<th>Query 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\alpha = 0$</td>
<td></td>
</tr>
<tr>
<td>The event is in [MASK].</td>
<td>progress</td>
<td>The event is in London.</td>
<td>London</td>
<td>The event is in August.</td>
</tr>
<tr>
<td></td>
<td>June</td>
<td></td>
<td>Dublin</td>
<td></td>
</tr>
<tr>
<td></td>
<td>July</td>
<td></td>
<td>Edinburgh</td>
<td></td>
</tr>
<tr>
<td></td>
<td>April</td>
<td></td>
<td>Paris</td>
<td></td>
</tr>
<tr>
<td></td>
<td>September</td>
<td></td>
<td>Sydney</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Toronto</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>London</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>June</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>July</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>March</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>September</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>April</td>
<td></td>
</tr>
<tr>
<td>The book is for [MASK].</td>
<td>children</td>
<td>The book is for him.</td>
<td>me</td>
<td></td>
</tr>
<tr>
<td></td>
<td>women</td>
<td></td>
<td>her</td>
<td>free</td>
</tr>
<tr>
<td></td>
<td>adults</td>
<td></td>
<td>him</td>
<td>sale</td>
</tr>
<tr>
<td></td>
<td>sale</td>
<td></td>
<td>you</td>
<td>download</td>
</tr>
<tr>
<td></td>
<td>boys</td>
<td></td>
<td>us</td>
<td>reading</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>free</td>
<td>children</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\times$</td>
<td></td>
</tr>
</tbody>
</table>

#29
Interpolation: Results

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

- **Novel methodology** and **dataset** for investigating the geometry of the BERT-space
  - interpretation of arbitrary points

- Conclusions about the BERT-space:
  - substantial regularity, with regions that correspond to distinct senses
  - evidence for “voids”—regions that do not correspond to any intelligible sense.

Limitations & Future Work:

- Short and carefully constructed sentences → naturalistic sentences
- English only → other languages
- Representations in ambiguous contexts?
THE END!