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

Motivating Example

ambiguous word = What is the sense?

The event is in _____

The event is in London

The event is in October

Motivating Example

The event is in London

locative sense

disambiguating token

The event is in October

temporal sense

disambiguating token

Motivating Example

The event is in _____

The event is in London

locative sense

disambiguating token

The event is in October

temporal sense

disambiguating token

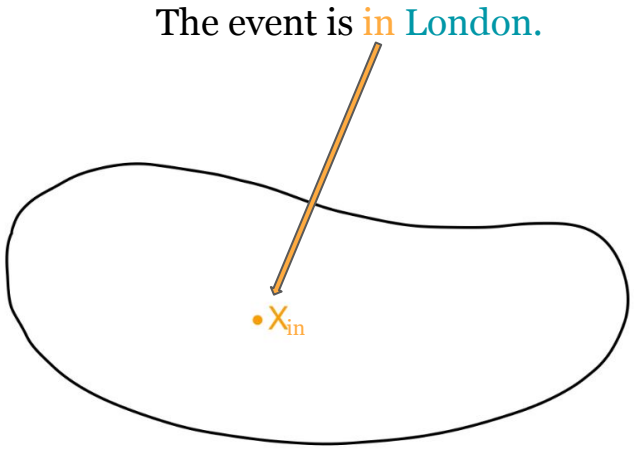
How is the information about the **sense** encoded in the contextualised representation of “in”?

Main Hypothesis

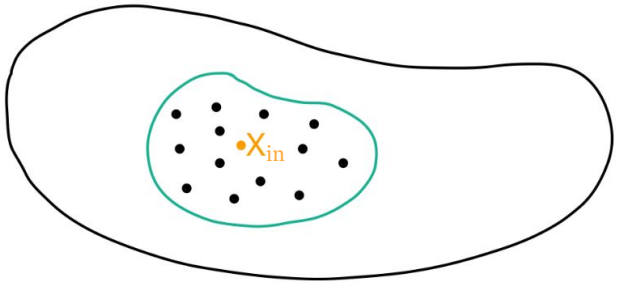
There are regular “nicely defined” regions in the BERT-space around words that correspond to distinct senses.

Geometric View

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



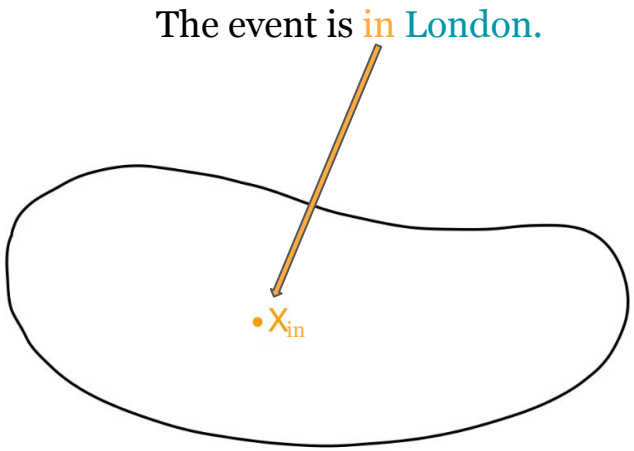
BERT-space



BERT-space

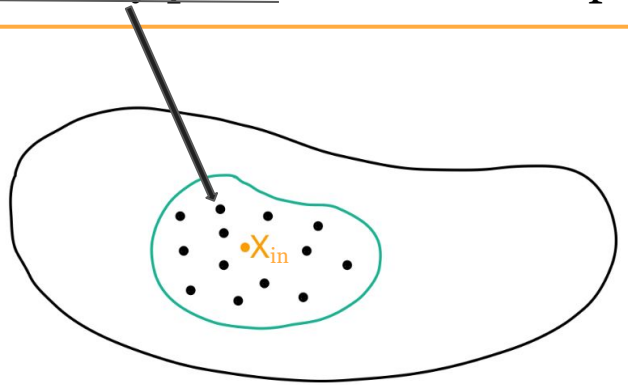
Geometric View

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



BERT-space

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



BERT-space

Masked Pseudoword Probing (MaPP)

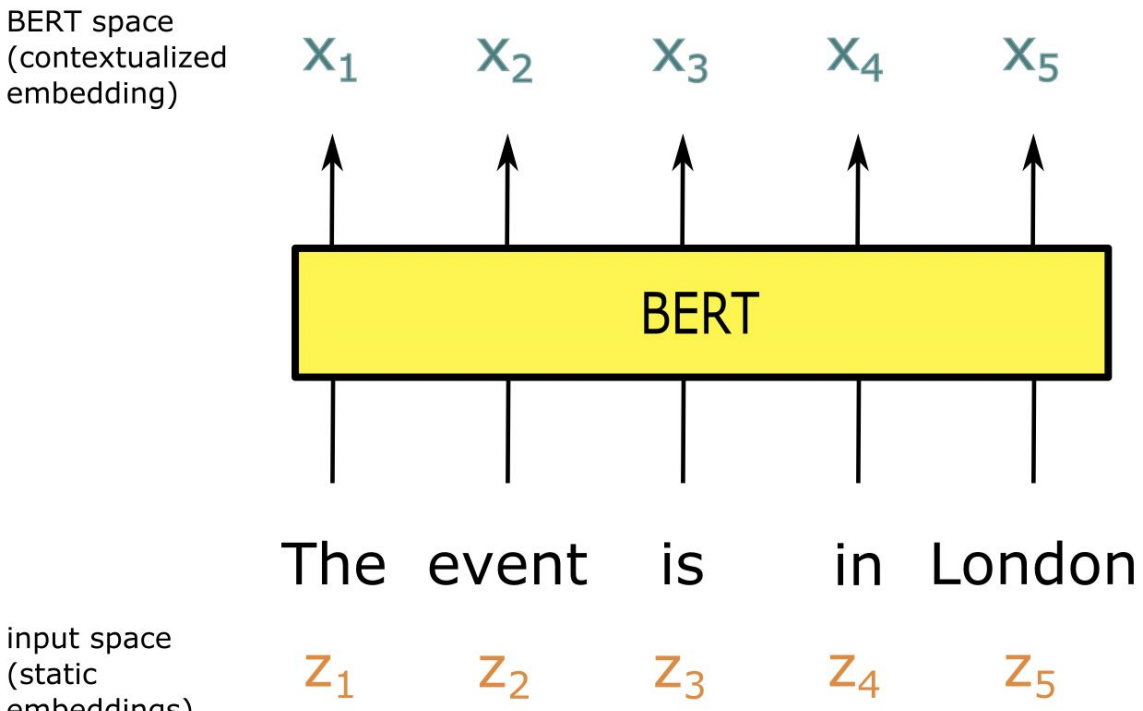
Novel technique to investigate the geometry of the BERT-space in a controlled manner around individual instances.

Masked Pseudoword Probing (MaPP)

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

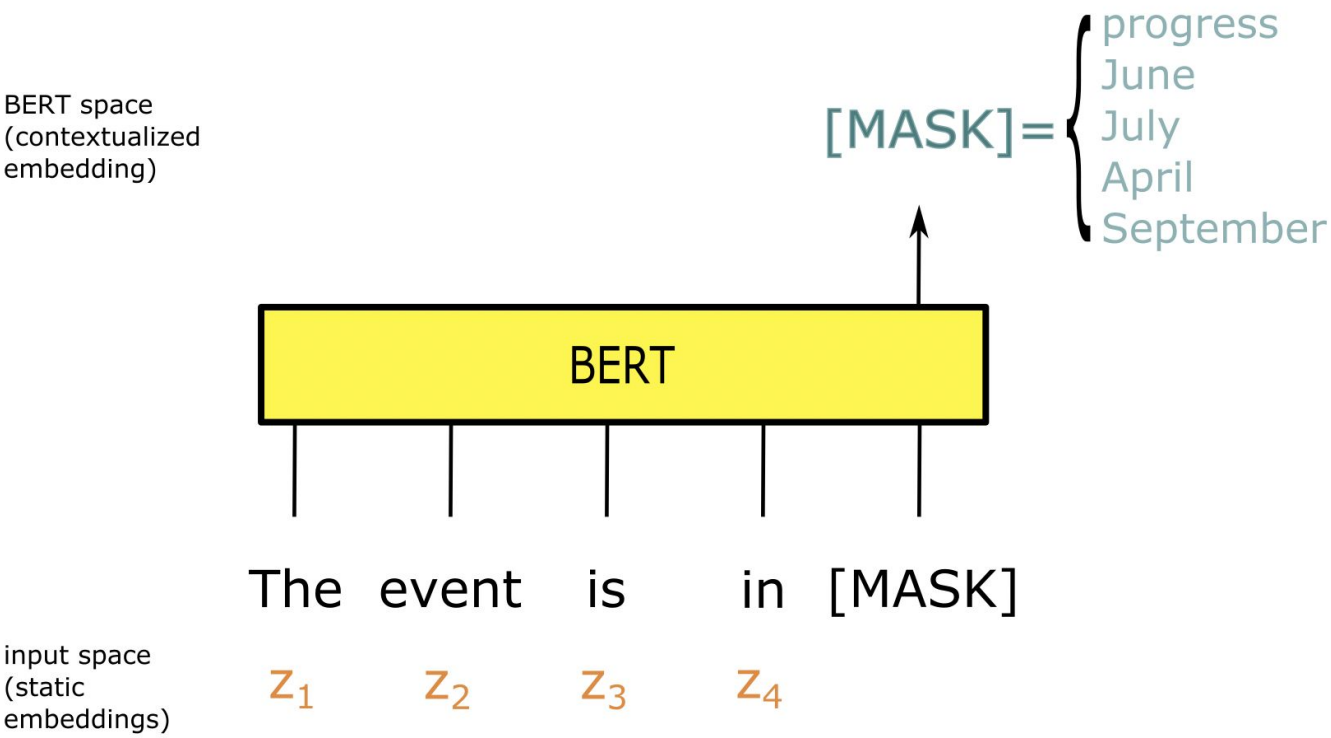
How do different regions in the contextualised space correspond to word senses?

Masked Pseudoword Probing (MaPP)

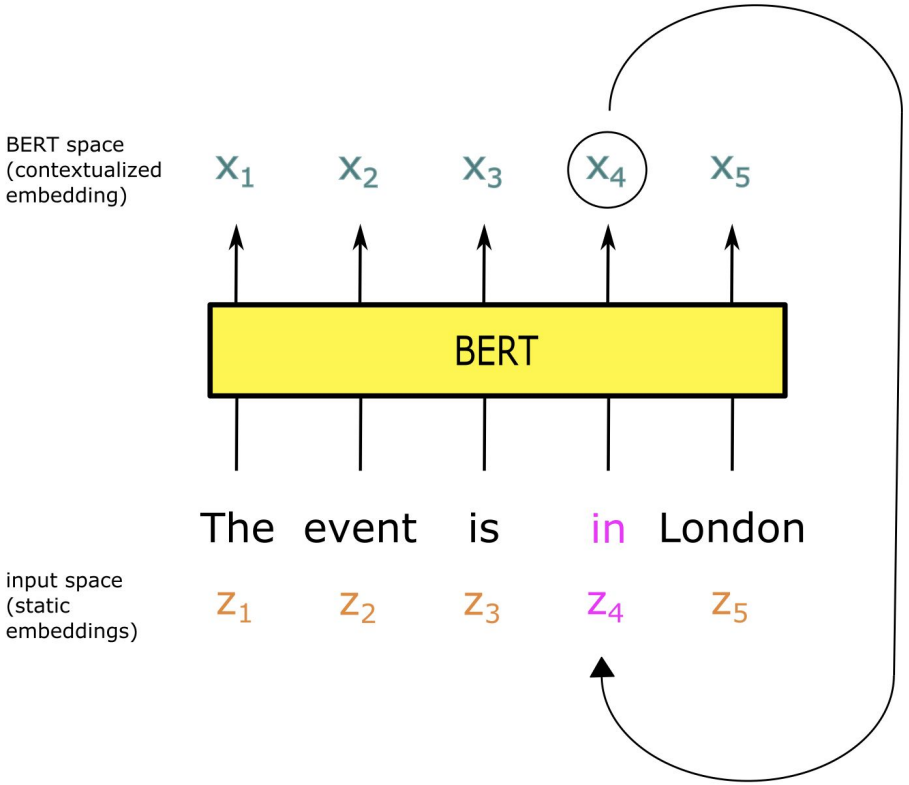


Masked Pseudoword Probing (MaPP)

We can also use BERT for masked prediction



Masked Pseudoword Probing (MaPP)



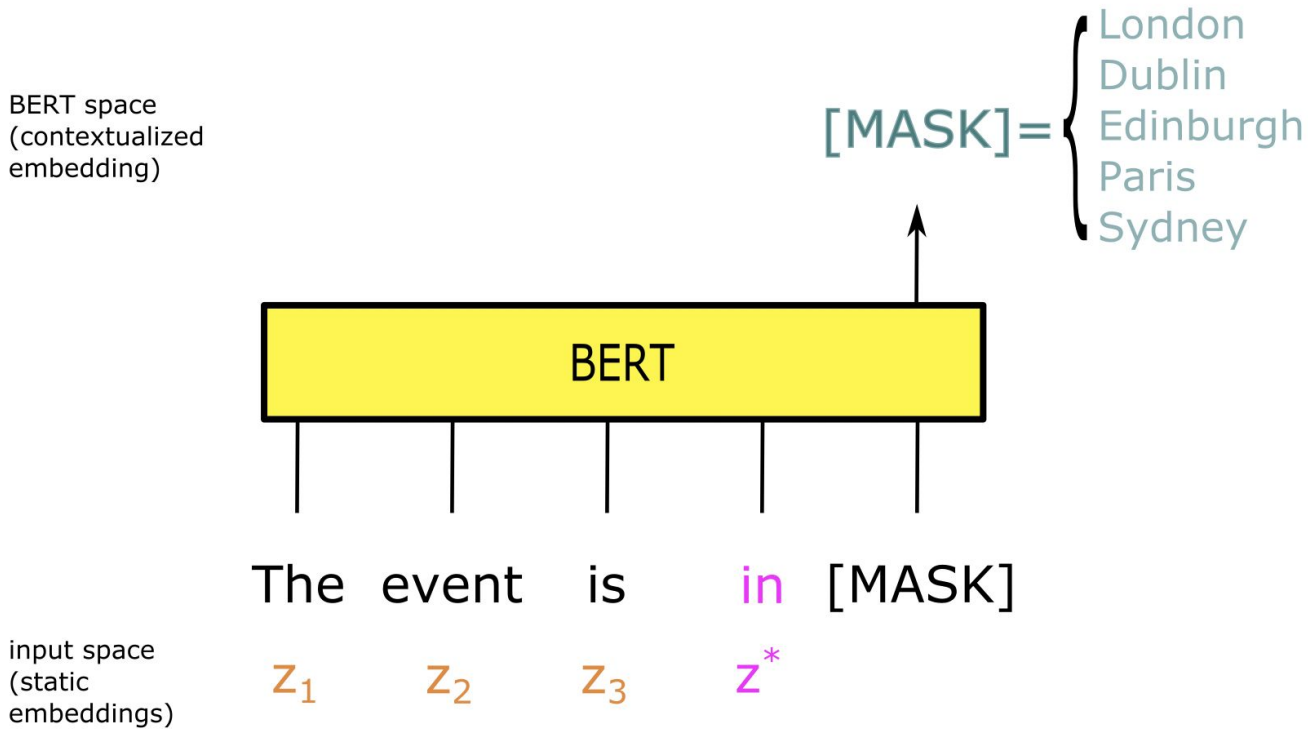
We learn a pseudoword $\mathbf{z}^*$ in place of z_4 which is customized to reconstruct x_4

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

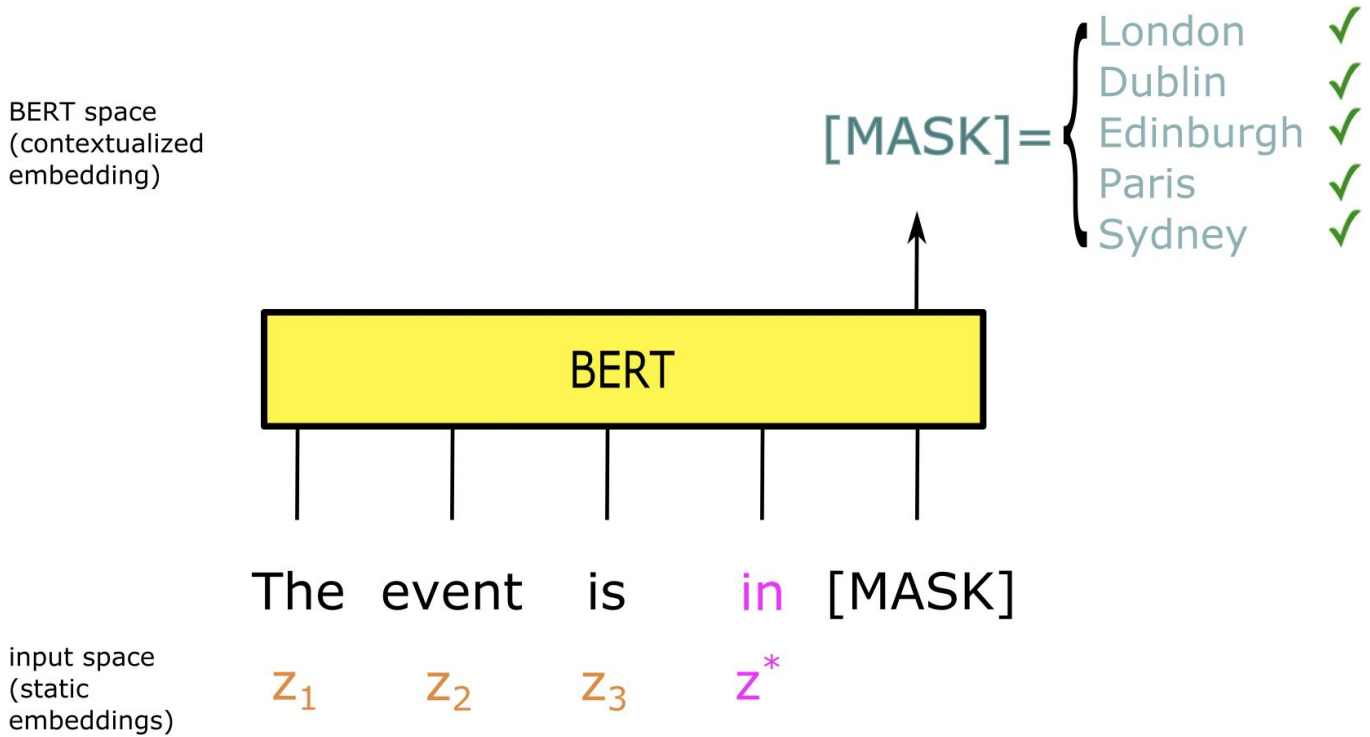
(here t=4, d=768)

Masked Pseudoword Probing (MaPP)

Masked prediction using a pseudoword

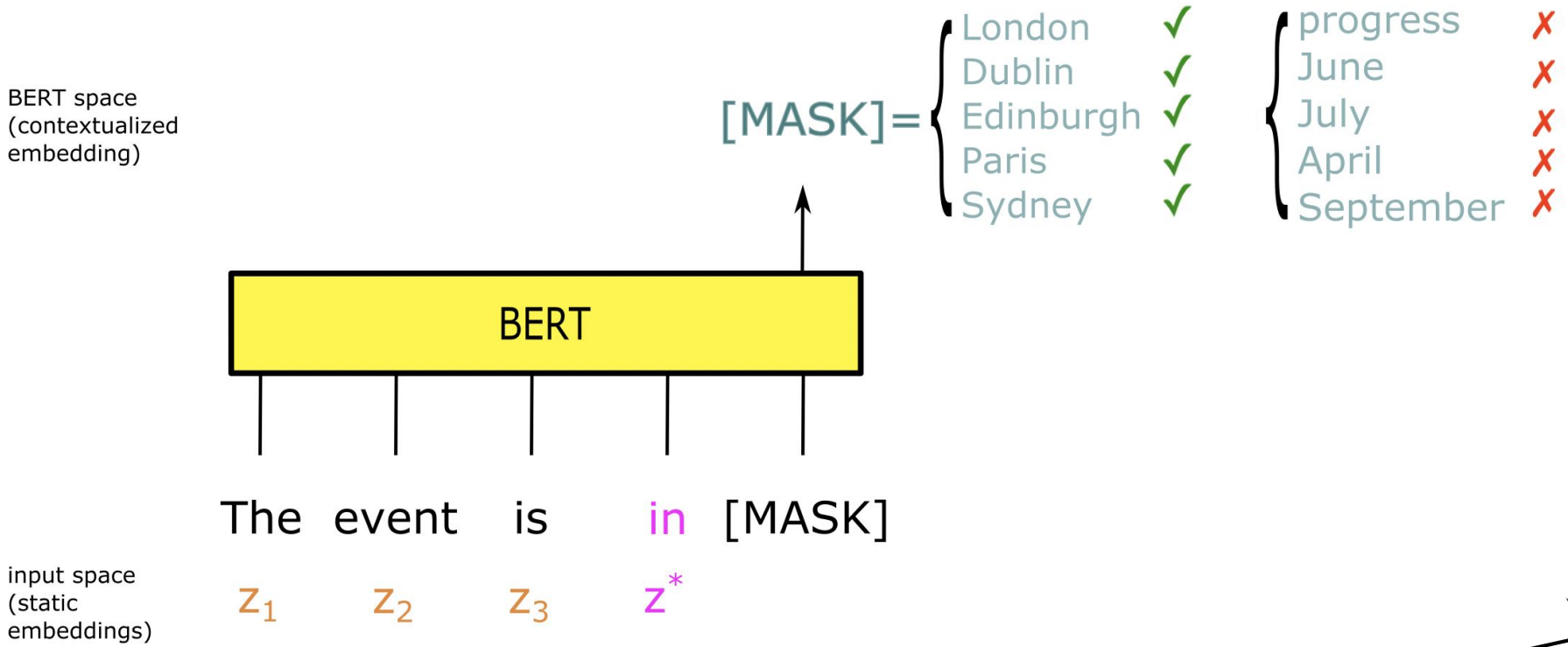


Masked Pseudoword Probing (MaPP)



Masked Pseudoword Probing (MaPP)

original static embedding



Masked Pseudoword Probing (MaPP): Summary

1. Run BERT for a sentence.

BERT(The event is in October.)

static inputs: $\mathbf{z}_{\text{The}}$ $\mathbf{z}_{\text{event}}$ $\mathbf{z}_{\text{is}}$ $\mathbf{z}_{\text{in}}$ ...
contextualized outputs: $\mathbf{x}_{\text{The}}$ $\mathbf{x}_{\text{event}}$ $\mathbf{x}_{\text{is}}$ $\mathbf{x}_{\text{in}}$ ...

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

2. Learn pseudoword $\mathbf{z}^*_{\text{in}}$ in place of $\mathbf{z}_{\text{in}}$ that is customized to reconstruct $\mathbf{x}_{\text{in}}$.

3. Run masked prediction with the modified input vector.

BERT(The event is **in** [MASK].)
$\mathbf{z}_{\text{The}}$ $\mathbf{z}_{\text{event}}$ $\mathbf{z}_{\text{is}}$ $\mathbf{z}^*_{\text{in}}$

4. Examine top predictions for sense match.

October ✓ July ✓ winter ✓
London ✗ ##ic ✗

decoding

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.

Focus Word	Sentence	Sense
in	The event is **in October**.	temporal
for	The book is **for Lisa**.	person
with	I ate salad **with enjoyment**.	feeling
about	The clip is **about** a **horse**.	topic
started	I **started** the **car**.	device
had	I **had** a **party**.	social event
had	I **had slept**.	auxiliary/past participle

Research Questions & Experiments

Experiment 1: **Specialization**

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

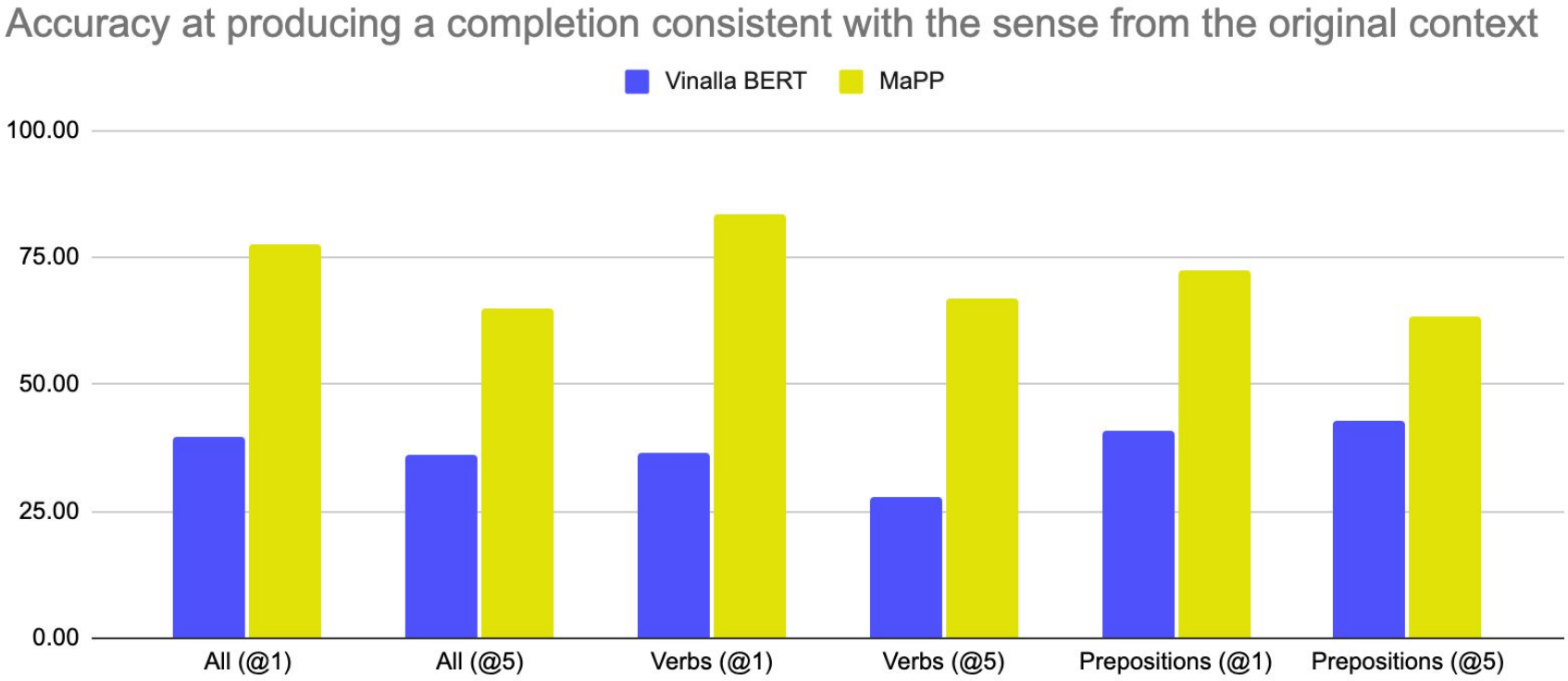
Research Questions & Experiments

Experiment 1: **Specialization**

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

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

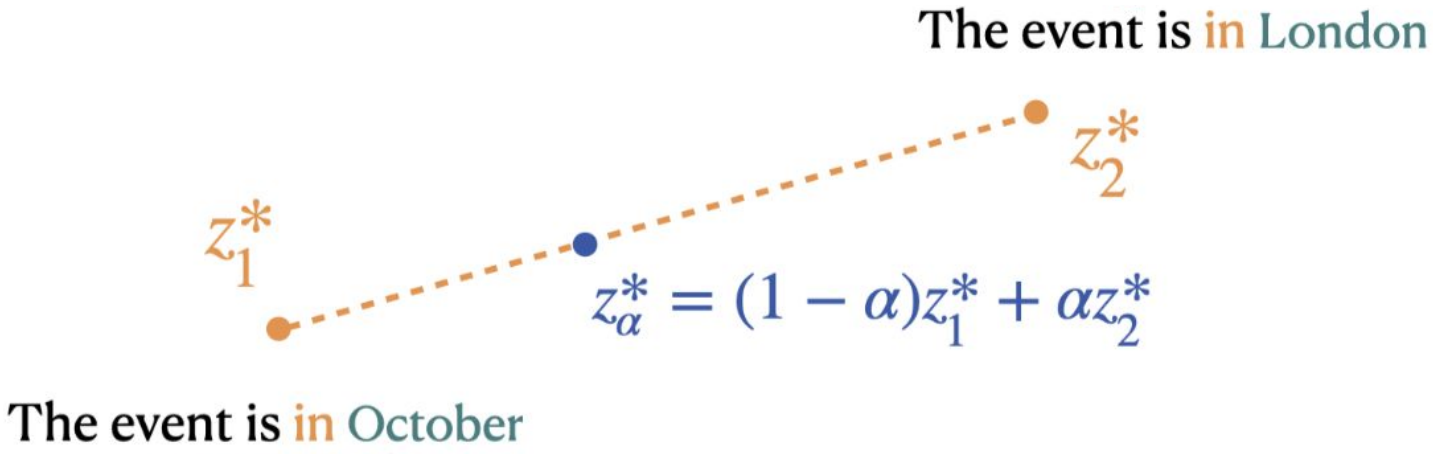
Specialization Experiment: Results



Research Questions & Experiments: Interpolation

Experiment 2: **Interpolation**

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



Examples for the interpolation results:

Mask	Vanilla BERT	Query 1	Interpolated MaPP				Query 2
			$\alpha = 0$	$\alpha = 0.4$	$\alpha = 0.8$	$\alpha = 1$	
The event is in [MASK].	progress ✗ June ▶ July ▶ April ▶ September ▶	The event is **in London**.	**London** ◀ Dublin ◀ Edinburgh ◀ Paris ◀ Sydney ◀	Toronto ◀ **London** ◀ June ◀ Dublin ◀ Melbourne ◀	June ▶ July ▶ March ▶ September ▶ April ▶	July ▶ September ▶ June ▶ March ▶ **August** ▶	The event is **in August**.
The book is for [MASK].	children ◀ women ◀ adults ◀ sale ▶ boys ◀	The book is **for him**.	me ◀ her ◀ **him** ◀ you ◀ us ◀	children ◀ women ◀ you ◀ sale ▶ free ✗	free ✗ sale ▶ download ▶ reading ▶ children ◀	free ✗ download ▶ sale ▶ reading ▶ purchase ▶	The book is **for viewing**.

Interpolation: Results

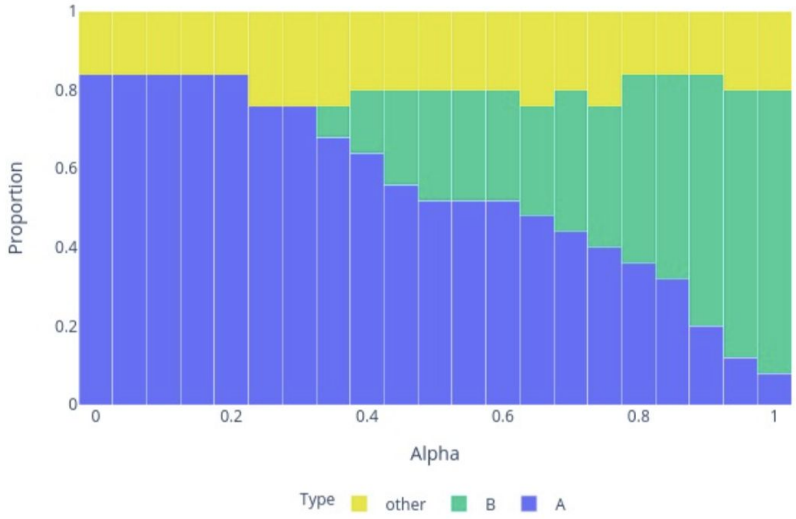


Figure 5: Interpolation results for minimal pair data as a function of interpolation parameter α: average proportion of top-1 predictions consistent with sense A, which predominates at $\alpha = 0$; sense B, which predominates at $\alpha = 1$; or neither.

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!