# SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings

EMNLP 2022

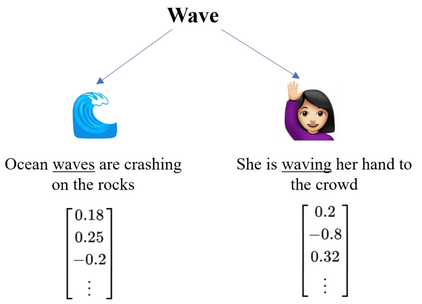
Jan Engler[1], Sandipan Sikdar[2], **Marlene Lutz**[3], Markus Strohmaier[3,4,5]

[1] RWTH Aachen University, [2] L3S Research Center, [3] University of Mannheim, [4] GESIS, [5] CSH Vienna

## INTRODUCTION

**Objective:** Adding interpretability to pre-trained contextual word embeddings

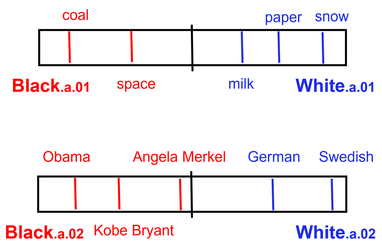
**Problem:** Words in different contexts have different representations




We need to deal with polysemy!

**Approach:**

- Transformation to a new embedding space with *interpretable* dimensions
- Dimensions are defined by polar opposite *word senses*



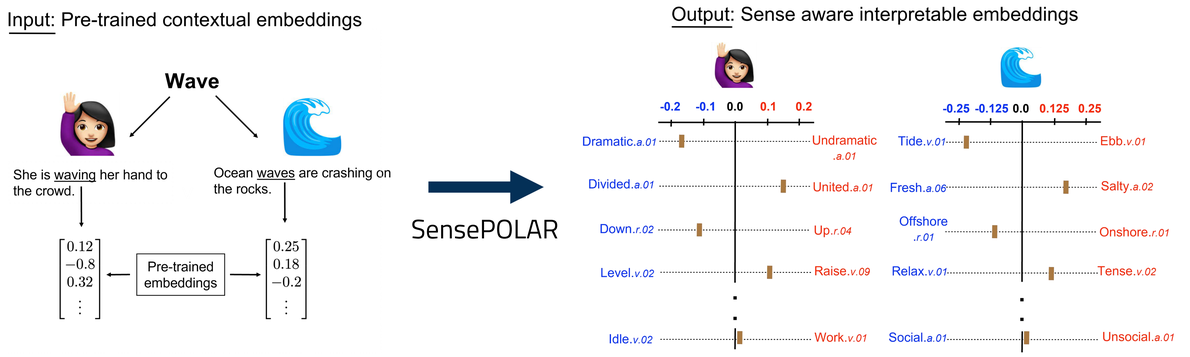

## CONTACT

Marlene Lutz

✉ marlene.lutz@uni-mannheim.de

@mar_lutz

## FRAMEWORK




## METHODOLOGY

### 1. ORACLE

Obtaining polar sense dimensions from an oracle

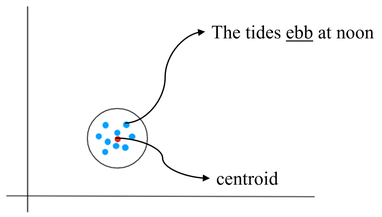
Oracle

Tide.v.02 - Ebb.v.01
Idle.v.02 - Work.v.01
⋮

Here: WordNet as an oracle

### 2. POLAR SENSE EMBEDDINGS

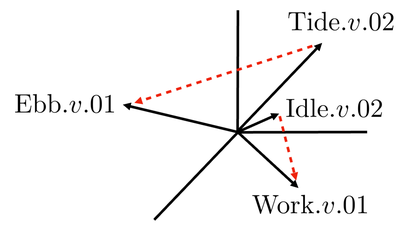
- Retrieving BERT embeddings of a word in sense-specific context examples
- Creating a *sense embedding* as the centroid of all sense-specific word embeddings




$$\text{Ebb}.v.01 = \frac{1}{m}\sum_{j=1}^{m} \text{Ebb}_{c_j}^{v.01}$$

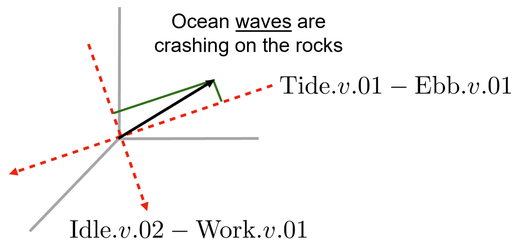
### 3. POLAR SENSE SPACE

Each new dimension is the difference between two polar opposite sense embeddings




### 4. TRANSFORMATION

Transforming any word embedding into the polar sense space by change of basis:




## EVALUATION

### FINE-TUNING TASKS

BERT vs SensePOLAR on the SQuAD benchmark

| | SQuAD 1.1 | | SQuAD 2.0 | |
|---|---|---|---|---|
| Metric | Base | SensePOLAR | Base | SensePOLAR |
| EM | 86.92 | 86.85↓ 0.07% | 80.88 | 81.06↑ 0.22% |
| F1 | 93.15 | 93.12↓ 0.03% | 83.87 | 83.89↑ 0.02% |

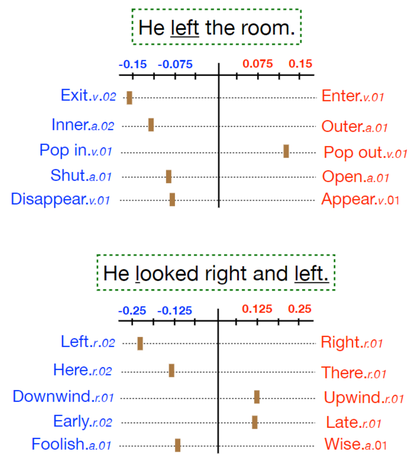
### SURVEY EXPERIMENT

Cond. probability of top dimensions selected by SensePOLAR to be chosen by human annotators

| Top-$k$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| SensePOLAR | 0.876 | 0.558 | 0.312 | 0.187 | 0.093 |
| Random | 0.5 | 0.22 | 0.083 | 0.023 | 0.004 |

### INTERPRETABILITY

SensePOLAR can differentiate between senses




Find us on GitHub:

/JanEnglerRWTH/SensePOLAR