# Correlations between sets of Words

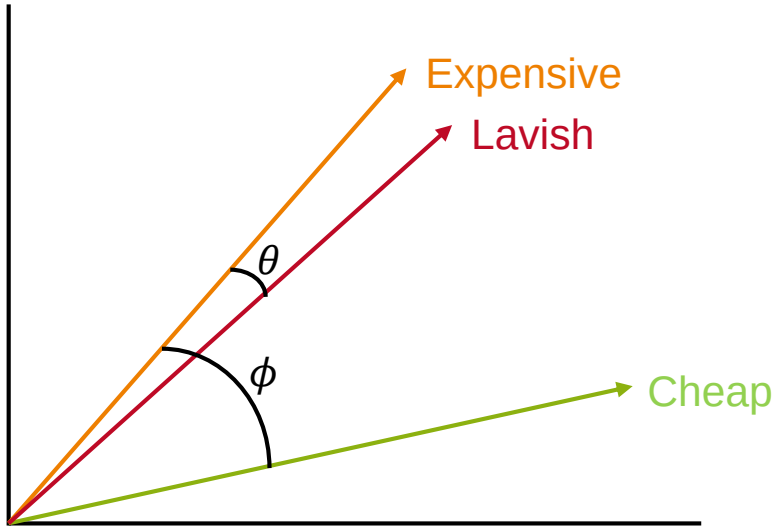
Carel van Niekerk

03.07.2020

# Contents

# Problem Description

- Given word embeddings for some words we can find semantic similarity between these words using cosine similarity.



# Problem Description

Heinrich Heine
Universität Düsseldorf

- Now we have two sets of words:

  Not to pricey

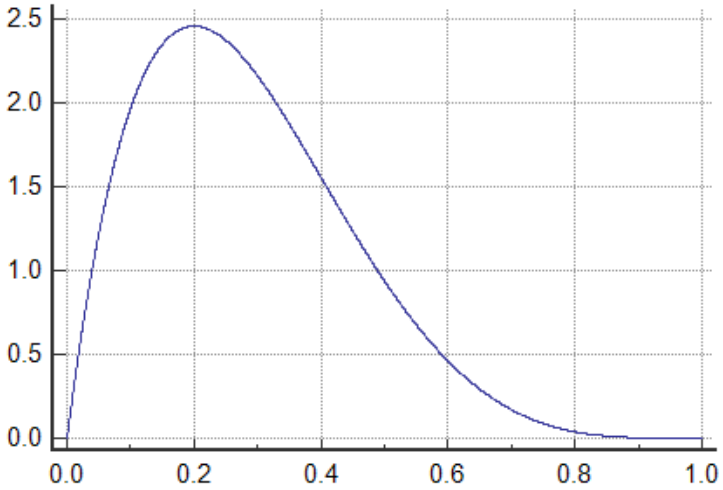
  Costing an arm and a leg
- Embedding models will now give us two sets of embeddings
- How do we find the similarity between these two sets?

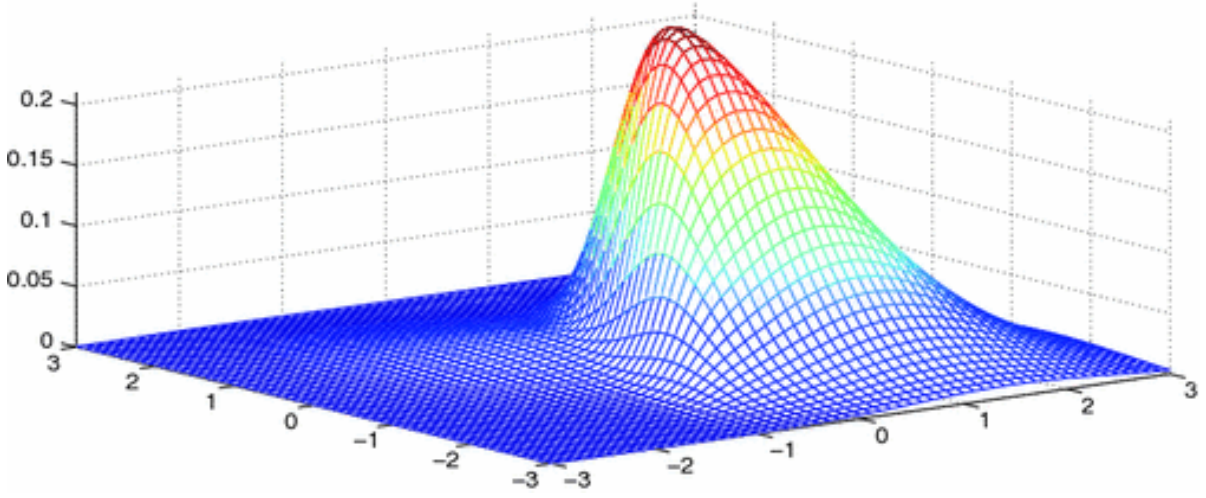
# Two different views on word distributions

Heinrich Heine
Universität Düsseldorf

1. Each word embedding comes from a ***d*-dimensional distribution**.

2. Each word has its **own semantic distribution** (1-dimensional). The word embedding of that word is a set of observations from this distribution.



# Distributions over words

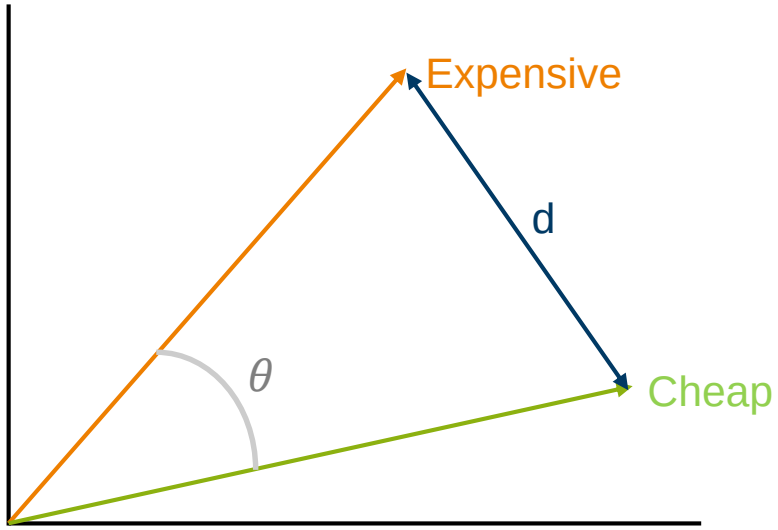
$\boldsymbol{W}_i \sim F_D(\cdot)$

- Each word embedding is an observation from the $d$-dimensional distribution over semantic meanings.

- Similarity Measures:
  - Cosine Similarity
  - Euclidean Distance

- Cannot use correlation since we do not know the true distribution and we only have 1 observation of each word so it cannot be estimated.

# Similarity



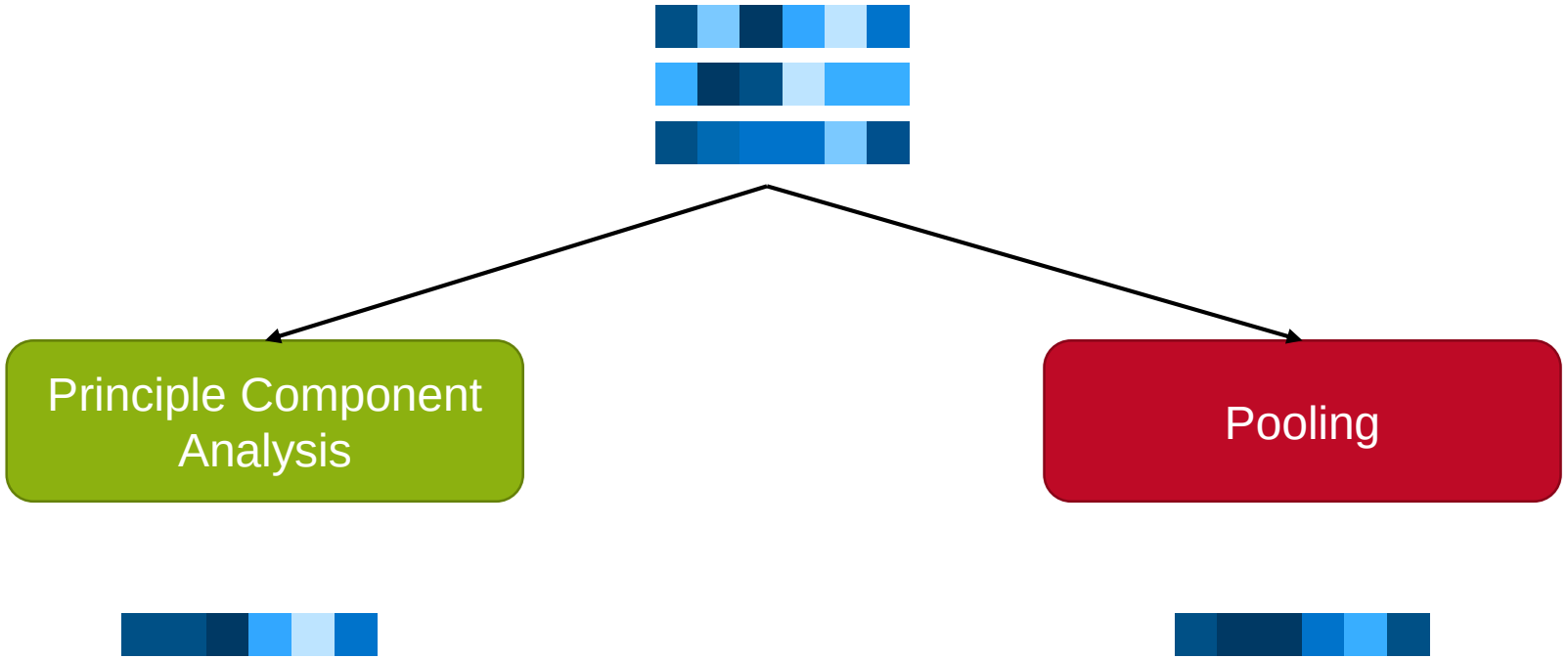

- Cosine Similarity

  Sim(Expensive, Cheap) = $\cos(\theta)$

- Euclidean Distance (L2 Norm)

  Sim(Expensive, Cheap) = d

# Similarity of sets of words



# Similarity of sets of words

Heinrich Heine
Universität Düsseldorf

| Approach | STS 12 | STS 13 | STS 14 | STS 15 | STS 16 |
|---|---|---|---|---|---|
| PCA | **58.6** | **67.3** | **70.5** | **73.5** | **71.7** |
| Mean Pool | 58.3 | 57.9 | 64.9 | 67.6 | 64.3 |
| Max Pool | 57.7 | 53.5 | 67.2 | 69.5 | 68.5 |

Correlation between human annotated similarity and predicted similarity


Correlations between Word Vector Sets. Zhelezniak et.al. 2019

# Distribution of a word

$W_i \sim F_i(\cdot)$

$\boldsymbol{w}_i \in \mathbb{R}^D$

- Each dimension in the word vector is an observation from the word distribution over semantic meanings.

- Similarity Measures:
  - Correlation Coefficient

# Similarity

Heinrich Heine
Universität Düsseldorf

- Correlation

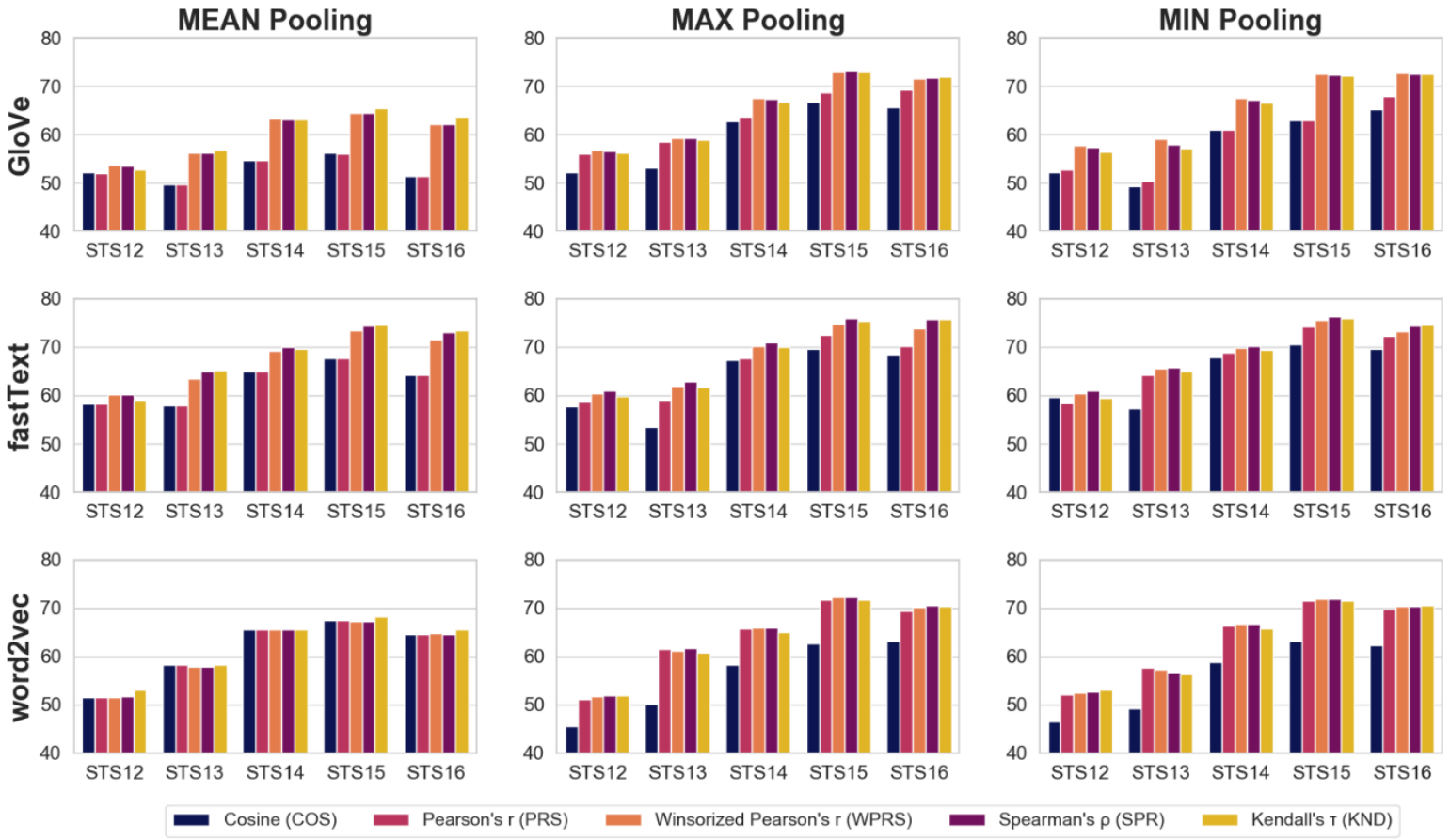
  Sim(Expensive, Cheap)

  $= Corr\left(W_{Expensive}, W_{Cheap}\right)$

  $\approx \frac{\hat{\sigma}_{Expensive,Cheap}}{\hat{\sigma}_{Expensive}\,\hat{\sigma}_{Cheap}}$ (Estimates based on the observed word vectors)

  If the average of the word vectors is 0, which is the case for models such as BERT, then we have that the correlation is equal to the cosine similarity.

# Results






Correlations between Word Vector Sets. Zhelezniak et.al. 2019

# Reproducing Kernel Hilbert Spaces (RKHS)

Heinrich Heine
Universität Düsseldorf

- Let **X** be an Arbitrary set and **H** a Reproducing Kernel Hilbert Space with kernel **K**.
- **H** is a space of function over **X**.
- Then for all x in **X** there exists a unique function $K_x$ in **H** with the reproducing property $f(x) = \langle f, K_x \rangle_H$ for all f in **H**.
- $\langle \cdot, \cdot \rangle_H$ is the inner product under **H**.

- It has been shown that RKHS can be used to find the distance between two sets using only their Kernel functions. This is due to the reproducing property.


https://www.youtube.com/watch?v=Cl4qrGaRJZo

# Correlations between sets of words

Heinrich Heine
Universität Düsseldorf

- Consider two sets of word embeddings $S_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$, $S_2 = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_M\}$
- We construct two RKHS $\mathcal{F}$ and $\mathcal{G}$ over these two sets.
- Then for kernel functions $K$ and $L$ we can calculate the Hilbert-Schmidt independence criterion (HSIC) which indicates the level of dependence between the two sets.
- For $\boldsymbol{K} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\mathbf{L} = L(\boldsymbol{y}_i, \boldsymbol{y}_j)$:

$$HSIC(S_1, S_2, K, L) = HSIC(\boldsymbol{K}, \boldsymbol{L})$$

# Correlations between sets of words

Heinrich Heine
Universität Düsseldorf

- A generalisation of the Pearson's correlation coefficient for these sets is the Centred Kernel Alignment (CKA). The CKA is defined as:

$$CKA(\boldsymbol{K}, \boldsymbol{L}) = \frac{HSIC(\boldsymbol{K},\boldsymbol{L})}{\sqrt{HSIC(\boldsymbol{K},\boldsymbol{K})HSIC(\boldsymbol{L},\boldsymbol{L})}}$$

- HSIC is an operator comparable to Covariance and hence CKA is comparable to the Pearson's correlation coefficient.

# Results

Heinrich Heine
Universität Düsseldorf

| Approach | STS 12 | STS 13 | STS 14 | STS 15 | STS 16 |
|---|---|---|---|---|---|
| PCA | 58.6 | **67.3** | 70.5 | 73.5 | 71.7 |
| MeanPool | 58.3 | 57.9 | 64.9 | 67.6 | 64.3 |
| MaxPool | 57.7 | 53.5 | 67.2 | 69.5 | 68.5 |
| Linear CKA | 59.8 | 62.1 | 69.5 | 74.6 | 70.3 |
| Gaussian CKA | 60.5 | 63.8 | **71.6** | **76.3** | **73.7** |
| dCor CKA | **61.0** | 63.2 | 71.5 | 75.6 | 72.4 |

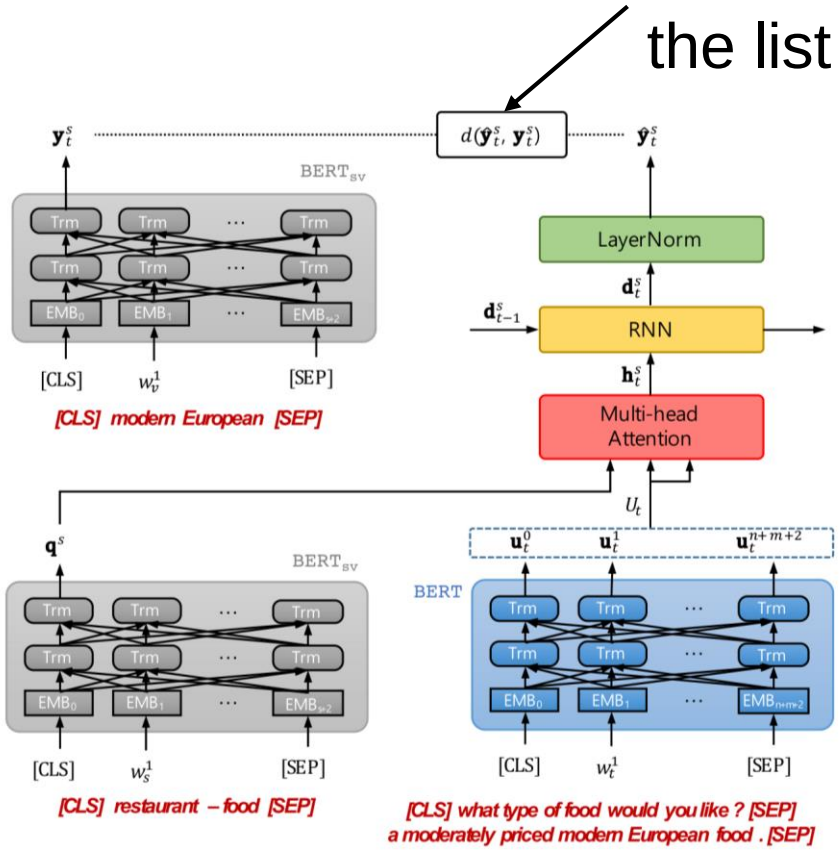
Correlation between human annotated similarity and predicted similarity

# Applications in Dialogue

Heinrich Heine
Universität Düsseldorf

- Semantic Similarity is a very useful tool in striving towards ontology independent Dialogue State Tracking.
  - Additional useful information and descriptions come in the form of sets of words.
- Semantic Similarity can be a useful tool in evaluating responses.
  - Reponses are also sets of words
- This correlation between sets of words measure can be a useful tool in dialogue modelling.

# Applications in Dialogue




- Currently a pooled approach with cosine similarity is used.
- This could potentially be replaced by the improved CKA correlation.

# Conclusion

Heinrich Heine
Universität Düsseldorf

- Statistical correlation between sets of words promises to be a better estimate of semantic similarity than cosine similarity based on pooled / PCA vectors.

- Good semantic similarity measures combined with good embedding models could possibly improve ontology independent dialogue models.

Heinrich Heine
Universität Düsseldorf

# Questions

# References


- A kernel method for the two-sample problem. Gretton et.al. 2007
- Correlations between Word Vector Sets. Zhelezniak et.al. 2019
- How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. Ethayaraijh 2019