

# A Probabilistic Approach to Sense Embeddings

Probabilistic Machine Learning Course Project

---

Anuj Nagpal, Divyat Mahajan and Rushab Munot

IIT Kanpur

# Table of contents

1. Problem Description
2. Previous Work
3. Approaches We Tried
4. Results

# Problem Description

---

# Problem Description

- The problem of generating word vectors is of utmost importance to any task in Natural Language Processing.

# Problem Description

- The problem of generating word vectors is of utmost importance to any task in Natural Language Processing.
- A one vector per word approach is inadequate to model polysemous words.

# Problem Description

- The problem of generating word vectors is of utmost importance to any task in Natural Language Processing.
- A one vector per word approach is inadequate to model polysemous words.
- A single word can have multiple meanings when used in different contexts.

# Problem Description

Consider two senses of the word bank:

# Problem Description

Consider two senses of the word bank:

- One pertaining to the financial sense,
- And the other to the bank of a river.



Consider two senses of the word bank:

- One pertaining to the financial sense,
- And the other to the bank of a river.

## Issues

These two senses of bank are hardly related to each other in any way, however both of them have the same vector, which is sort of a weighted combination of the two senses.

# Problem Description

Consider two senses of the word bank:

- One pertaining to the financial sense,
- And the other to the bank of a river.

## Issues

These two senses of bank are hardly related to each other in any way, however both of them have the same vector, which is sort of a weighted combination of the two senses.

This distortion is unwanted. Thus, we need a better model that can learn multiple vectors per word, every vector corresponding to the sense of a word.

## Previous Work

---

# Multimodal Word Distributions

- Athiwaratkun and Wilson proposed a probabilistic word embedding that can capture multiple meanings.

# Multimodal Word Distributions

- Athiwaratkun and Wilson proposed a probabilistic word embedding that can capture multiple meanings.
- They represented each word  $w$  as a Gaussian mixture with  $K$  components such that the density function of  $w$  is given by:

$$p_w(\vec{x}) = \sum_{i=1}^K p_{w,i} \mathcal{N}[\vec{x}; \vec{\mu}_{w,i}, \Sigma_{w,i}]$$

$$= \sum_{i=1}^K \frac{p_{w,i}}{\sqrt{2\pi|\Sigma_{w,i}|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}_{w,i})^T \Sigma_{w,i}^{-1}(\vec{x}-\vec{\mu}_{w,i})}$$

where  $\sum_{i=1}^K p_{w,i} = 1$ .

They used a max-margin ranking objective used for Gaussian embeddings which pushes the similarity of a word and its positive context higher than that of its negative context by a margin  $m$ :

$$L_{\theta}(w, c, c') = \max(0, m - \log E_{\theta}(w, c) + \log E_{\theta}(w, c'))$$

- Neelakantan et. al. proposed a method based on context clustering for sense vector generation.

# Efficient Non-Parametric Estimation of Multiple Embeddings

- Neelakantan et. al. proposed a method based on context clustering for sense vector generation.
- They maintain a global vector for each word as well as multiple sense vectors and the context is embedded as a sum of the global vectors of the words in the context



# Efficient Non-Parametric Estimation of Multiple Embeddings

- They cluster these average contexts, increasing the number of clusters every time a dissimilar context is observed, thus learning the number of senses per word ( a non parametric approach).

# Efficient Non-Parametric Estimation of Multiple Embeddings

- They cluster these average contexts, increasing the number of clusters every time a dissimilar context is observed, thus learning the number of senses per word ( a non parametric approach).
- This is essentially a Word Sense Disambiguation layer which is added before the skipgram layer in word2vec.

# Efficient Non-Parametric Estimation of Multiple Embeddings

- They cluster these average contexts, increasing the number of clusters every time a dissimilar context is observed, thus learning the number of senses per word ( a non parametric approach).
- This is essentially a Word Sense Disambiguation layer which is added before the skipgram layer in word2vec.
- Thus a disambiguated sense vector is used to predict the context in the skip gram layer.

## Approaches We Tried

---

# Approach 1

- What if we propose an approach where instead of learning a separate Gaussian for each sense of a word, we learn a set of basis Gaussian vectors  $\{(\mu_1, \Sigma_1), (\mu_2, \Sigma_2), \dots, (\mu_m, \Sigma_m)\}$ ?

# Approach 1

- What if we propose an approach where instead of learning a separate Gaussian for each sense of a word, we learn a set of basis Gaussian vectors  $\{(\mu_1, \Sigma_1), (\mu_2, \Sigma_2), \dots, (\mu_m, \Sigma_m)\}$ ?
- Using this set of basis Gaussian vectors, we now model the  $k$ th sense of word  $w$  by:  
$$w_{i,k} = \sum_{m=0}^M z_{i,k,m} * \mathcal{N}(\mu_m, \Sigma_m)$$
 where  $z_{i,k,m}$  serve as latent variables for each  $w$  to be learned.

# Approach 1

- For the  $i^{th}$  word,  $w_i$ , we incorporate all the senses as

$$w_i = \sum_{k=1}^K \pi_{i,k} * \sum_{m=1}^M z_{i,k,m} \mathcal{N}(\mu_m, \Sigma_m)$$

- For the  $i^{th}$  word,  $w_i$ , we incorporate all the senses as

$$w_i = \sum_{k=1}^K \pi_{i,k} * \sum_{m=1}^M Z_{i,k,m} \mathcal{N}(\mu_m, \Sigma_m)$$

- If  $\psi_{i,m} = \sum_{k=1}^K \pi_{i,k} * Z_{i,k,m}$



# Approach 1

- For the  $i^{th}$  word,  $w_i$ , we incorporate all the senses as

$$w_i = \sum_{k=1}^K \pi_{i,k} * \sum_{m=1}^M Z_{i,k,m} \mathcal{N}(\mu_m, \Sigma_m)$$

- If  $\psi_{i,m} = \sum_{k=1}^K \pi_{i,k} * Z_{i,k,m}$
- Then  $w_i = \sum_{m=1}^M \psi_{i,m} * \mathcal{N}(\mu_m, \Sigma_m)$

## Expected Improvement

- This model has total parameters to be learned as  $MD + MD^2 + NKM + NK$  whereas the previous one had  $KND + KND^2 + NK$

## Expected Improvement

- This model has total parameters to be learned as  $MD + MD^2 + NKM + NK$  whereas the previous one had  $KND + KND^2 + NK$
- With the following approximate rough values:  
 $N = 10^5, K = 5, D = 150, M = 100$ , there will be roughly 100 times reduction in number of parameters

- To actually compare senses across words there must be a specific relation between  $\Pi_i$  and  $Z_i$  for all words  $w_i$ .

- To actually compare senses across words there must be a specific relation between  $\Pi_i$  and  $Z_i$  for all words  $w_i$ .
- However, multiple solutions exist for the equations described in the previous slide and we can't maintain an invariant property.

## Approach 2

- What if we follow the same model as done by Athiwaratkun and Wilson but express each  $\mu_{w,k}$  as  $\sum_{m=1}^M z_{w,k,m} \mu_m$  and each  $\Sigma_{w,k}$  as  $\sum_{m=1}^M z_{w,k,m} \Sigma_m$  where  $\mu_m$  and  $\Sigma_m$  are shared by all words?

## Results

---

# Parameters for Original Approach

- K 2
- D 50



# Parameters for Approach 1

- K 2
- M 10
- D 50

## Parameters for Approach 2

- K 2
- M 300
- D 50

Approach	Average Precision	F1 Score
Original	67.726	72.442
1*	50.00	57.001
2	61.66	67.550

## Top 10 highest similarity

'islam:0', 'besht:1', 'prevail:1', 'shirk:0', 'sadducees:1', 'judaizers:1',  
'teachings:1', 'persecutions:1', 'belief:0', 'conversion:0'

## Top 10 highest similarity

'islam:0', 'besht:1', 'prevail:1', 'shirk:0', 'sadducees:1', 'judaizers:1',  
'teachings:1', 'persecutions:1', 'belief:0', 'conversion:0'

## Top 10 lowest variance of top 20 highest similarity

'sadducees:1', 'tabari:1', 'persecutions:1', 'halakhic:0', 'sages:0',  
'salafis:1', 'judaizers:1', 'shirk:0', 'samaritan:1', 'sects:0', 'besht:1',  
'scripture:1', 'atheism:1', 'prevail:1', 'teachings:1', 'islam:0', 'yehuda:1',  
'belief:0', 'conversion:0', 'hebrew:0'

## Top 10 highest similarity

'islam:1', 'islamic:1', 'muslim:1', 'muslims:0', 'sunni:0', 'shia:1',  
'muhammad:0', 'arab:1', 'religious:0', 'wahhab:0'

## Top 10 highest similarity

'islam:1', 'islamic:1', 'muslim:1', 'muslims:0', 'sunni:0', 'shia:1',  
'muhammad:0', 'arab:1', 'religious:0', 'wahhab:0'

## Top 10 lowest variance of top 20 highest similarity

'jihad:1', 'sunni:0', 'shia:1', 'wahhab:0', 'muhammad:0', 'sects:0',  
'brotherhood:1', 'sharia:1', 'muslims:0', 'arabs:1', 'faiths:1', 'sect:0',  
'druze:1', 'islam:1', 'muslim:1', 'islamic:1', 'arab:1', 'bah:0', 'religion:0',  
'religious:0'

### Top 10 highest similarity

islam:0 baptist:1 fathers:0 canonical:1 judaism:1 religions:1 eusebius:1  
baptism:1 circumcision:1 anglican:1



### Top 10 highest similarity

islam:0 baptist:1 fathers:0 canonical:1 judaism:1 religions:1 eusebius:1  
baptism:1 circumcision:1 anglican:1

### Top 10 lowest variance of top 20 highest similarity

circumcision:1 canonical:1 baptism:1 liturgies:0 eusebius:1 canon:1  
religions:1 christians:1 scriptures:1 baptist:1 rabbis:0 doctrine:0  
fathers:0 apostolic:0 orthodoxy:1 anglican:1 judaism:1 doctrines:0  
islam:0 theologian:1

### Top 10 highest similarity

islam:1 muslim:1 arabs:0 arab:0 sunni:1 arabia:1 islamic:1 arabic:0  
tribal:1 syria:0

### Top 10 highest similarity

islam:1 muslim:1 arabs:0 arab:0 sunni:1 arabia:1 islamic:1 arabic:0  
tribal:1 syria:0

### Top 10 lowest variance of top 20 highest similarity

empires:0 balkans:1 descendants:1 macedonia:1 tribes:1  
macedonian:1 arab:0 arabs:0 muslim:1 homeland:0 basques:1 arabic:0  
tribal:1 islamic:1 arabia:1 islam:1 europeans:1 sunni:1 syria:0 slavs:1

Thanks!

---