Modelling Text

GloVe Embeddings

What is GloVe

Main Ideas:

- Word occurrences are primary (available) statistics to unsupervised methods to learn word representations

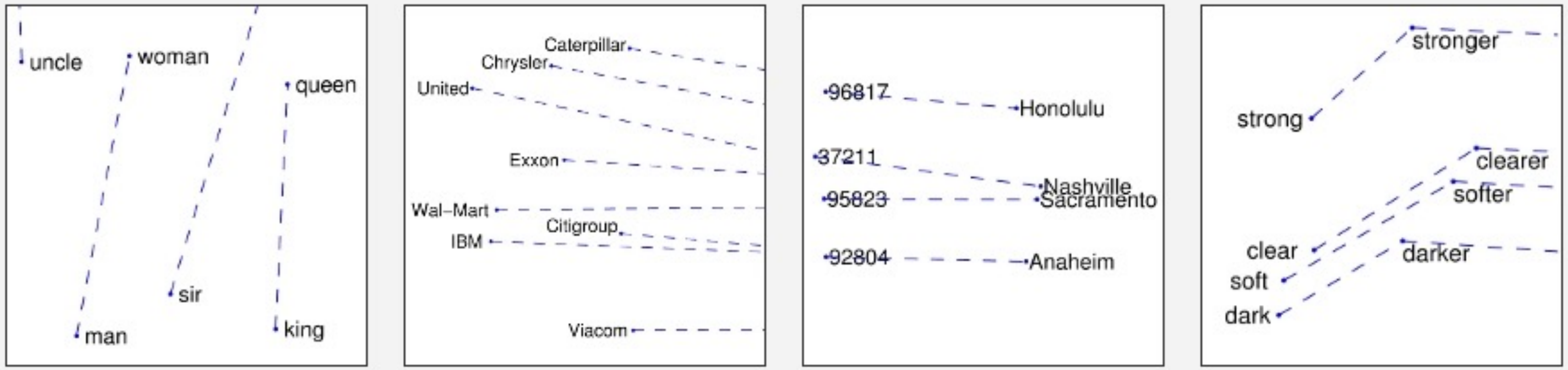
https://nlp.stanford.edu/pubs/glove.pdf

What is GloVe

- …baby don’t hurt me (https://nlp.stanford.edu/projects/glove/)

Introduction

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.



man - woman | company - ceo | city - zip code | comparative - superlative

GloVe -Basics

- X :   "Word-Word co-occurrence counts"
- X_{ij} : "frequency of word j occurring in context of i"
- $X_i = \sum_k X_{ik}$:   "frequency of any word in context of i"
- $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$: "probability that word j appears in the context of word i"

GloVe -Intuition

- We want to learn a model which predicts ratios instead of raw probabilities:

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

GloVe -Modelling

- Start with a general assumption:
  - We search a function F as follows:

$$F\left(w_i, w_j, \widetilde{w_k}\right) = \frac{P_{ik}}{P_{jk}}$$

Word vectors

Context word

GloVe -Modelling

- Start with a general assumption:
  - We search a function F as follows:

$$F\left(w_i, w_j, \widetilde{w_k}\right) = \frac{P_{ik}}{P_{jk}}$$

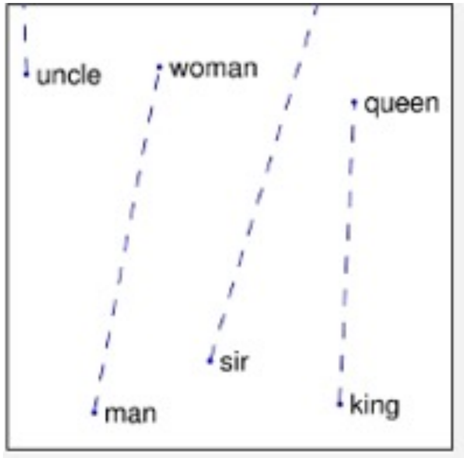
  - We could now try all possible functions for F and select the best one!
    - ➔ Search space is huge!
    - ➔ Iteratively integrate intuitive elements

GloVe -Modelling

- We search a function F as follows:

$$F(w_i, w_j, \widetilde{w_k}) = \frac{P_{ik}}{P_{jk}}$$

- Assumption 1:
  - The ratios should reflect similarities in the vector space



- Approach:

$$F(w_i, w_j, \widetilde{w_k}) = \frac{P_{ik}}{P_{jk}} \rightarrow F(w_i - w_j, \widetilde{w_k}) = \frac{P_{ik}}{P_{jk}}$$

GloVe -Modelling

- We search a function F as follows:

$$F\left(w_i, w_j, \widetilde{w_k}\right) = \frac{P_{ik}}{P_{jk}}$$

- Assumption 2:
  - Input are vectors while the output is a scalar

- Approach:
  - Take the dot product of the vectors

$$F\left(w_i - w_j, \widetilde{w_k}\right) = \frac{P_{ik}}{P_{jk}} \rightarrow F\left(\left(w_i - w_j\right)^T \cdot \widetilde{w_k}\right) = \frac{P_{ik}}{P_{jk}}$$

GloVe -Modelling

- We search a function F as follows:

$$F\left(w_i, w_j, \widetilde{w_k}\right) = \frac{P_{ik}}{P_{jk}}$$

- Assumption 3:
  - Words and their context words should be interchangeable

- Approach:
  - F has to be a group homomorphism between ($\mathbb{R}$,+) and ($\mathbb{R}$,x):

➔ one solution that worked is as follows:

$$F\left(\left(w_i - w_j\right)^T \cdot \widetilde{w_k}\right) = \frac{F(w_i^T \, \widetilde{w_k})}{F(w_j^T \, \widetilde{w_k})} = \frac{P_{ik}}{P_{jk}}$$

GloVe -Modelling

- Approach:
  - F has to be a group homomorphism between $(R, +)$ and (R, x):
- Group homomorphism:
- A function F is a group homomorphism if:

$$F(x * y) = F(x) \cdot F(y)$$

- Okay so we got:

$$F\left(\left(w_i - w_j\right)^T \cdot \widetilde{w_k}\right) = F(w_i^T \widetilde{w_k} + (-w_j^T\, \widetilde{w_k}))$$

➔ We got a homomorphism between “+” and “x” in $\mathbb{R}$

GloVe -Modelling

- Okay so we got:

$$F\left(\left(w_i - w_j\right)^T \cdot \widetilde{w_k}\right) = F\left(w_i^T \widetilde{w_k} + \left(-w_j^T \widetilde{w_k}\right)\right)$$

$$F\left(w_i^T \widetilde{w_k} + \left(-w_j^T \widetilde{w_k}\right)\right) = F\left(w_i^T \widetilde{w_k}\right) \cdot F\left(-w_j^T \widetilde{w_k}\right)$$

- Additionally for group homomorphism it holds:

$$\boxed{F(x^{-1}) = F(x)^{-1}}$$

$$\Rightarrow F\left(-w_j^T \widetilde{w_k}\right) = \frac{1}{F(w_j^T \widetilde{w_k})}$$

$$\Rightarrow F\left(\left(w_i - w_j\right)^T \cdot \widetilde{w_k}\right) = \frac{F(w_i^T \widetilde{w_k})}{F(w_j^T \widetilde{w_k})} = \frac{P_{ik}}{P_{jk}}$$

GloVe -Modelling

- We search a function F as follows:

$$F\left(w_i, w_j, \widetilde{w_k}\right) = \frac{P_{ik}}{P_{jk}}$$

- Now find a function which works for:

$$F\left(\left(w_i - w_j\right)^T \cdot \widetilde{w_k}\right) = \frac{F\left(w_i^T\ \widetilde{w_k}\right)}{F\left(w_j^T\ \widetilde{w_k}\right)}$$

→ F(…) = exp(…)

Optimize with
Gradient Descent
For whole
vocabulary V

$$\exp\left(w_i^T\ \widetilde{w_k}\right) = P_{ik}$$

$$\boxed{w_i^T\ \widetilde{w_k} = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)}$$

GloVe –Full story

- Optimize this with a Linear regression objective:

$$J = \sum_{i=1}^{V}\sum_{j=1}^{V} f(X_{ij})\left(w_i^T w_j + b_i + b_j \; - \log(X_{ij})\right)^2$$

word scaling, based on α

$\log(X_i)$

Integrated to have an equal amount of parameters for every word

Summary

- Distributional (vector) models of meaning
  - **Sparse** (PPMI-weighted word-word co-occurrence matrices)
  - **Dense**:
    - Word-word SVD 50-2000 dimensions
    - Skip-grams and CBOW
    - Brown clusters 5-20 binary dimensions.
    - GloVE: State of the art word embeddings!