

# GloVe: Global Vectors for Word Representation

Damyana Gateva

**Saarland University**

November 30, 2017



# Overview

- 1 Introduction  
Motivation
- 2 The GloVe model
- 3 Experiments  
Training  
Evaluation
- 4 Results
- 5 Model Analysis
- 6 Conclusion

# Outline

## 1 Introduction

- Motivation

## 2 The GloVe model

## 3 Experiments

- Training
- Evaluation





## 4 Results

## 5 Model Analysis

## 6 Conclusion

# Motivation

- Nearest Neighbors: scalar distance between vectors

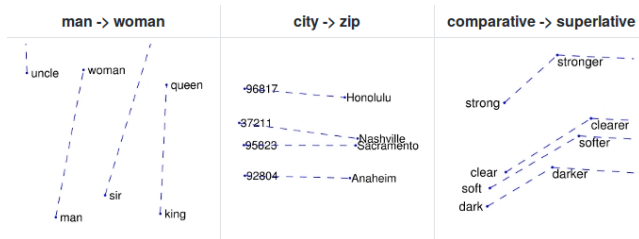
Litoria	Leptodactylidae	Rana	Eleutherodactylus
			

COO

# Motivation

- Linear substructures: various dimensions of difference
- Word analogies: test for linear relationships  $a:b::c:?$

$$\text{king} - \text{queen} = \text{man} - \text{woman}$$



# Introduction

- Existing Methods
  - ▶ Global Matrix factorization methods (LSA):
    - + use efficiently statistical information
    - perform bad on word analogy tasks
  - ▶ Local context window methods (word2vec skipgram):
    - + perform good on word analogy tasks
    - scan each context window in the corpusdo not use statistics from it directly
- Goal: create a model that combines the positive sides

# Outline

- 1 Introduction
  - Motivation

- 2 The GloVe model

- 3 Experiments
  - Training
  - Evaluation

- 4 Results

- 5 Model Analysis

- 6 Conclusion

# The GloVe model

- co-occurrence matrix  $X_{ij}$      $i$  center word,  $j$  context word

$$P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$$

- instead of probabilities: **probability ratio**

$$\frac{P_{ijk}}{P_{jk}}$$

$i = \text{ice}, j = \text{steam}, k = \text{solid}$



# The GloVe model

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

# Encoding Meaning in vector differences

Ratios of co-occurrence probabilities can encode meaning

- Looking for an encoding function  $F$ :  $F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$

# Encoding Meaning in vector differences

## Ratios of co-occurrence probabilities can encode meaning

- Looking for an encoding function  $F$ :  $F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$
- Log-bilinear model:  $w_i w_j = \log P(i|j)$
- Vector differences:  $w_x(w_a - w_b) = \log \frac{P(x|a)}{P(x|b)}$

# Encoding Meaning in vector differences

## Ratios of co-occurrence probabilities can encode meaning

- Looking for an encoding function  $F$ :  $F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$
- Log-bilinear model:  $w_i w_j = \log P(i|j)$
- Vector differences:  $w_x(w_a - w_b) = \log \frac{P(x|a)}{P(x|b)}$
- For each word pair:  $\log(X_{ik}) = w_i^T \tilde{w}_k + b_i + \tilde{b}_k$

## Least squares regression model with cost function J

$$J = \sum_{i,j=1}^V f(X_{ij})(w_i^T \tilde{w}_k + b_i + \tilde{b}_k - \log(X_{ik}))^2$$

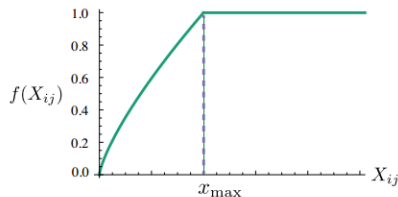
# Least squares regression model with cost function J

$$J = \sum_{i,j=1}^V f(X_{ij})(w_i^T \tilde{w}_k + b_i + \tilde{b}_k - \log(X_{ik}))^2$$

$$f(x) = \begin{cases} (x/x_{max})^\alpha & \text{if } x \leq x_{max} \\ 1 & \text{otherwise} \end{cases}$$

$$x_{max} = 100$$

$$\alpha = 3/4$$



# Outline

- 1 Introduction
  - Motivation
- 2 The GloVe model
- 3 Experiments**
  - Training
  - Evaluation
- 4 Results
- 5 Model Analysis
- 6 Conclusion

# Training

- Corpora
  - ▶ Wikipedia 2010 - 1 billion tokens
  - ▶ Wikipedia 2014 - 1.6 billion tokens
  - ▶ Gigaword5 - 4.3 billion tokens
  - ▶ Gigaword5 + Wikipedia 2014 - 6 billion tokens
  - ▶ Common Crawl - 42 billion tokens



# Training

- Constructing X

- ▶ filter vocabulary: 400,000 most frequent words (2 million for Common Crawl)
- ▶ choosing a context window size (default=10)
- ▶ symmetric/asymmetric: distinguishing left and right context or not
- ▶ decreasing weighting function:  
word pairs that are  $d$  words apart contribute  $1/d$  to the total count

# Training

- Training
  - ▶ train the model using stochastic gradient descent (AdaGrad)
  - ▶ sample non-zero elements from  $X$
  - ▶ iterations: 50 for vectors smaller than 300 dimensions, 100 otherwise
  - ▶ initial learning rate: 0.05
- Output:
  - ▶ 2 sets of word vectors:  $W$  and  $\tilde{W}$
  - ▶ resulting word vectors:  $W + \tilde{W}$

# Evaluation Methods

- Word analogies
- Word similarity
- Named entity recognition

# Outline

- 1 Introduction
  - Motivation
- 2 The GloVe model
- 3 Experiments
  - Training
  - Evaluation
- 4 Results**
- 5 Model Analysis
- 6 Conclusion

# Results: Word analogies

Model	Dim.	Size	Sem.	Syn.	Tot.
GloVe	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
GloVe	300	6B	80.8	61.5	70.3
SG	300	6B	73.0	66.0	69.1
CBOW	300	6B	63.6	67.4	65.7
GloVe	300	6B	77.4	67.0	71.7
GloVe	300	42B	<b>81.9</b>	<b>69.3</b>	<b>75.0</b>

## Results: Word similarity

Model	Size	WS353	MC	RG	SCWS	RW
SVD	6B	35.3	35.1	42.5	38.3	25.6
SVD-S	6B	56.5	71.5	71.0	53.6	34.7
SVD-L	6B	65.7	<u>72.7</u>	75.1	56.5	37.0
CBOW <sup>†</sup>	6B	57.2	65.6	68.2	57.0	32.5
SG <sup>†</sup>	6B	62.8	65.2	69.7	<u>58.1</u>	37.2
GloVe	6B	<u>65.8</u>	<u>72.7</u>	<u>77.8</u>	53.9	<u>38.1</u>
SVD-L	42B	74.0	76.4	74.1	58.3	39.9
GloVe	42B	<b><u>75.9</u></b>	<b><u>83.6</u></b>	<b><u>82.9</u></b>	<b><u>59.6</u></b>	<b><u>47.8</u></b>
CBOW*	100B	68.4	79.6	75.4	59.4	45.5

Spearman rank correlation on word similarity tasks

## Results: NER

Model	Dev	Test	ACE	MUC7
Discrete	91.0	85.4	77.4	73.4
SVD	90.8	85.7	77.3	73.7
CBOW	93.1	88.2	82.2	81.1
HPCA	92.6	<b>88.7</b>	81.7	80.7
GloVe	<b>93.2</b>	88.3	<b>82.9</b>	<b>82.2</b>

F1 score on NER task with 50d vectors

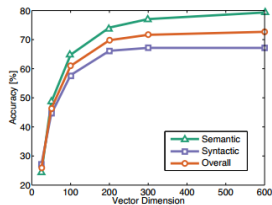
*Discrete* is the baseline without word vectors.

# Outline

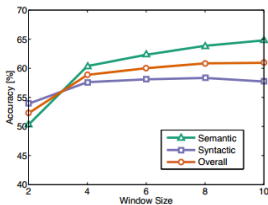
- 1 Introduction
  - Motivation
- 2 The GloVe model
- 3 Experiments
  - Training
  - Evaluation
- 4 Results
- 5 Model Analysis**
- 6 Conclusion



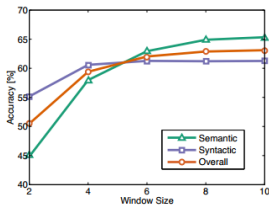
# Model Analysis: Vector Length and Context Size



(a) Symmetric context



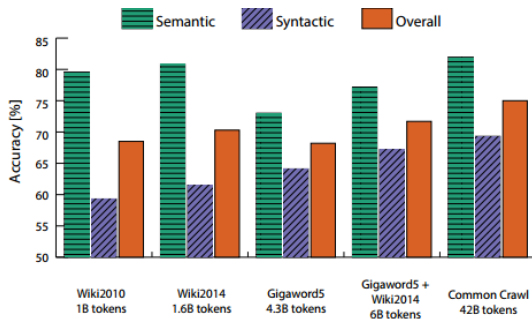
(b) Symmetric context



(c) Asymmetric context

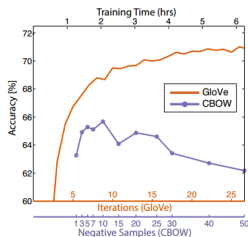
Accuracy on the analogy task as a function of vector size and window size/type

# Model Analysis: Corpus Size

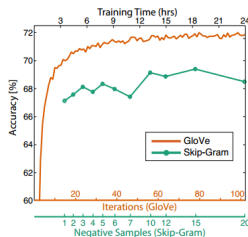


Accuracy on the analogy task for 300-dimensional vectors

# Model Analysis: Comparison with word2vec



(a) GloVe vs CBOW



(b) GloVe vs Skip-Gram

Accuracy on the word analogy task for 300-dimensional vectors as a function of training time

# Outline

- 1 Introduction
  - Motivation
- 2 The GloVe model
- 3 Experiments
  - Training
  - Evaluation
- 4 Results
- 5 Model Analysis
- 6 Conclusion

# Conclusion

- Prediction-based and count-based methods perform similarly
- Count-based methods capture global statistics more efficiently
- GloVe outperforms other models on word analogy, word similarity and NER tasks
- Fast training, scalable to huge corpora
- Good performance, even with small corpus and small vectors