NLP Basics
Corpus-Based Representation

國立臺灣大學 資訊工程學系
陳縕儂 助理教授
http://vivianchen.idv.tw

Meaning Representations in Computers

Knowledge-Based Representation

Corpus-Based Representation

Corpus-Based Representation

- Atomic symbols: ***one-hot*** representation

car $[0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ \ldots\ 0]$

car

Issues: difficult to compute the similarity (i.e. comparing “car” and “motorcycle”)

$[0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ \ldots\ 0]$ AND $[0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ \ldots\ 0] = 0$

car motorcycle

Idea: words with similar meanings often have similar neighbors

Corpus-Based Representation

- Neighbor-based representation
  - Co-occurrence matrix constructed via neighbors
  - Neighbor definition: full document v.s. windows

full document
word-document co-occurrence matrix gives general topics → “Latent Semantic Analysis”

windows
context window for each word → capture syntactic (e.g. POS) and sematic information

Window-Based Co-occurrence Matrix

- Example
  - Window length=1
  - Left or right context
  - Corpus:

I love AI.
I love deep learning.
I enjoy learning.

similarity > 0

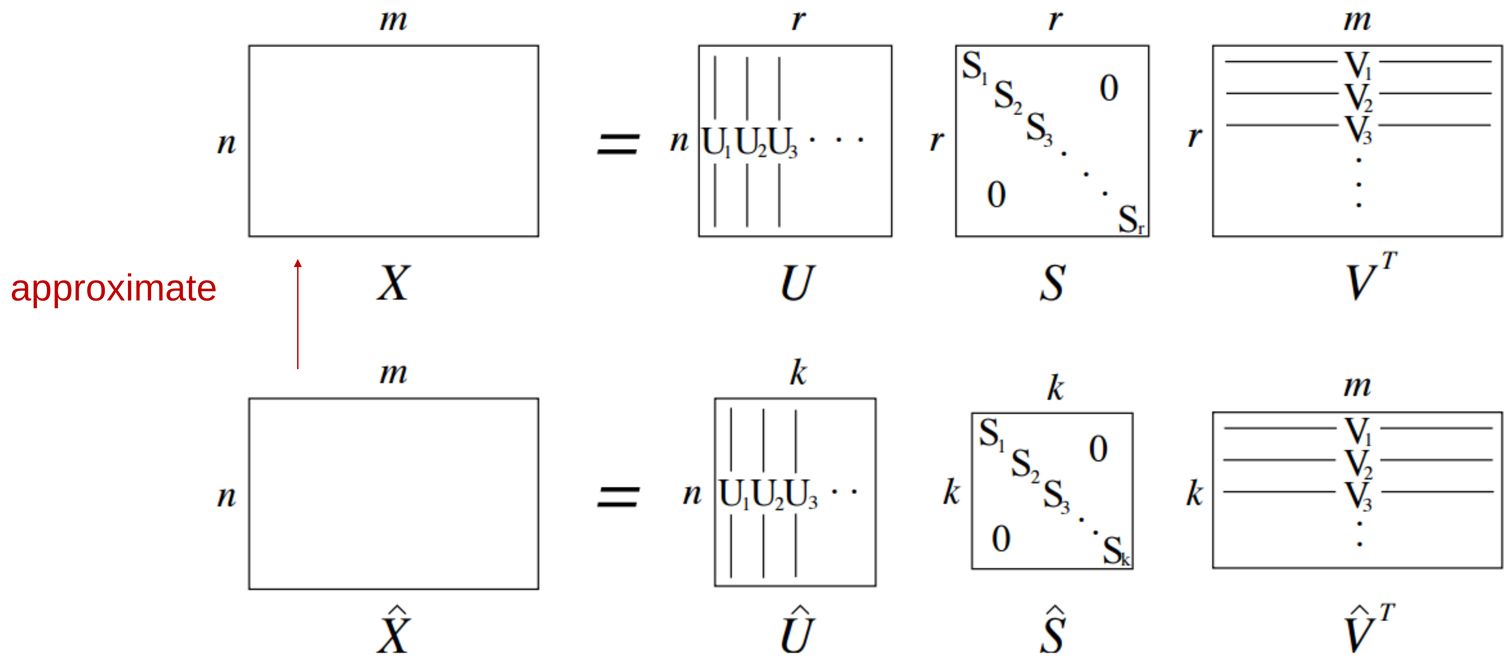
Counts	I	love	enjoy	AI	deep	learning
I	0	2	1	0	0	0
love	2	0	0	1	1	0
enjoy	1	0	0	0	0	1
AI	0	1	0	0	0	0
deep	0	1	0	0	0	1
learning	0	0	1	0	1	0

Issues:
- matrix size increases with vocabulary
- high dimensional
- sparsity → poor robustness

Idea: low dimensional word vector

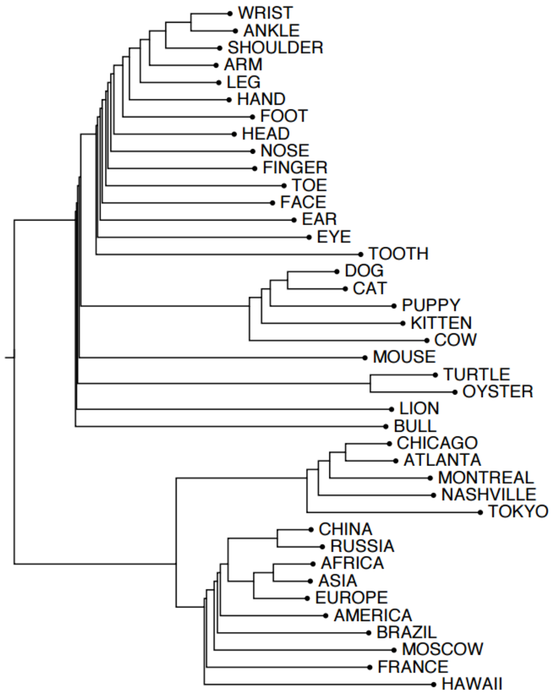
Low-Dimensional Dense Word Vector

- Method 1: dimension reduction on the matrix
- Singular Value Decomposition (SVD) of co-occurrence matrix X

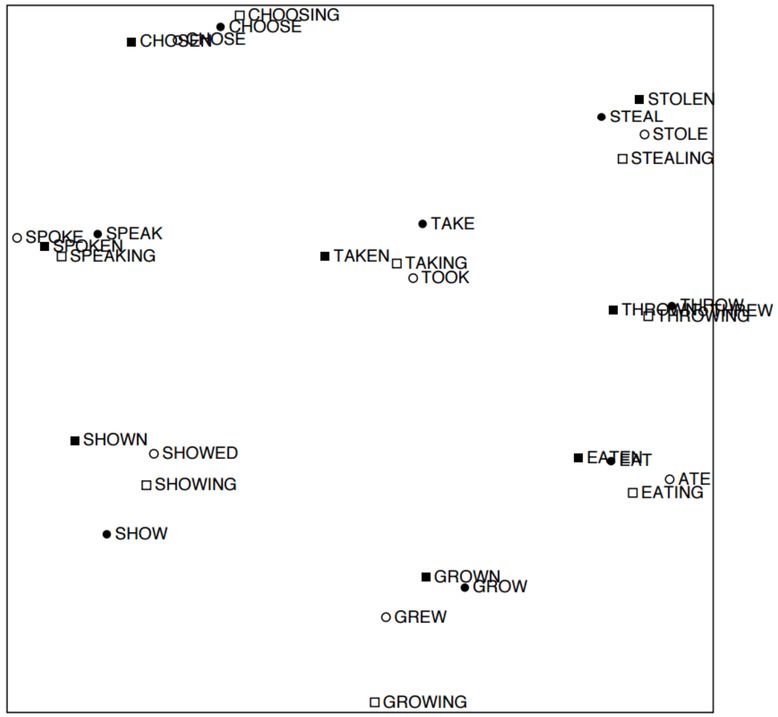


Low-Dimensional Dense Word Vector

- Method 1: dimension reduction on the matrix
- Singular Value Decomposition (SVD) of co-occurrence matrix X



semantic relations



syntactic relations

Issues:

- computationally expensive: $O(mn^2)$ when $n<m$ for nxm matrix
- difficult to add new words

Idea: directly learn low-dimensional word vectors

Rohde et al., “An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence,” 2005.

Low-Dimensional Dense Word Vector

- Method 2: directly learn low-dimensional word vectors
  - Learning representations by back-propagation. (Rumelhart et al., 1986)
  - A neural probabilistic language model (Bengio et al., 2003)
  - NLP (almost) from Scratch (Collobert & Weston, 2008)
  - Recent and most popular models: word2vec (Mikolov et al. 2013) and Glove (Pennington et al., 2014)
    - As known as “Word Embeddings”

Summary

- Knowledge-based representation
- Corpus-based representation
  - ✓ Atomic symbol
  - ✓ Neighbors
    - ○ High-dimensional sparse word vector
    - ○ Low-dimensional dense word vector
      - ▪ Method 1 – dimension reduction
      - ▪ Method 2 – direct learning