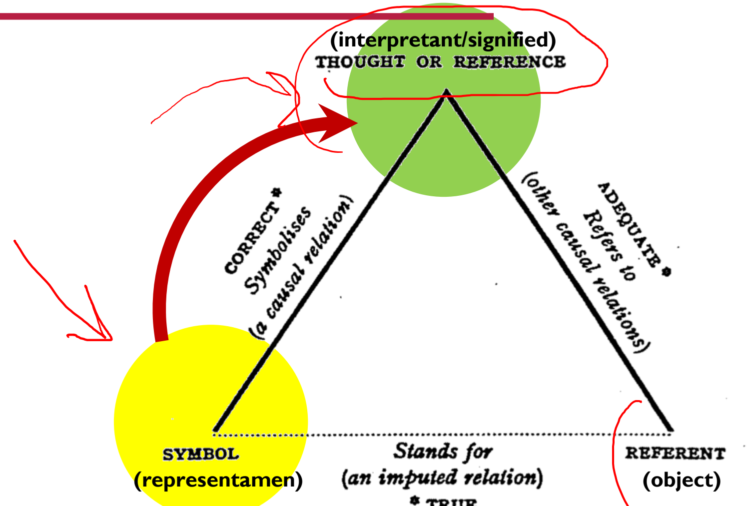

Ceci n'est pas une pipe.
Magritte

Representament → Interpretant



Ludwig Josef Johann Wittgenstein

/ˈvɪtɡənʃtaɪn, -staɪn/

1889 –1951

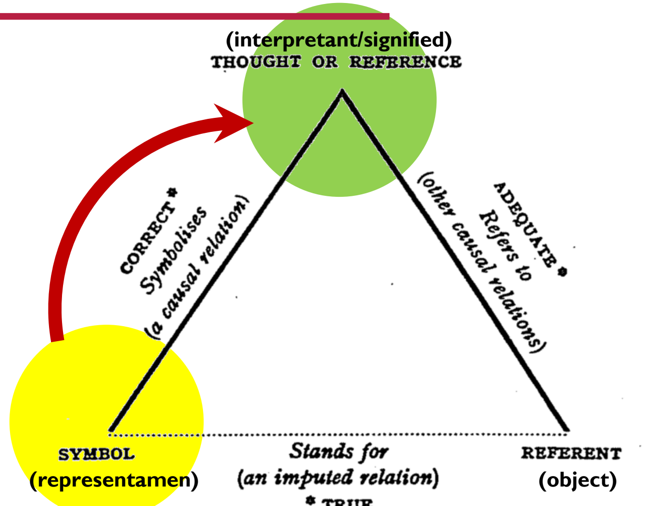
Austrian-British Philosopher

ʃtɔp → stop

Skeptical of a completely formal theory of meaning definitions for each word

⟶ "the meaning of a word is its use in the language" - Philosophical Investigations.

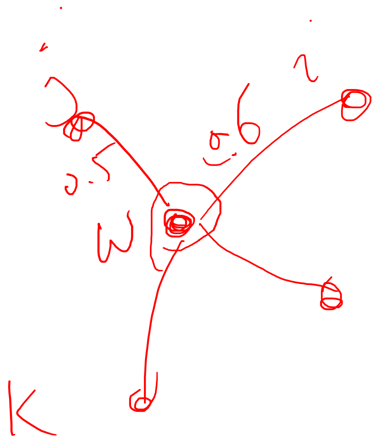
Token → Relations with other Tokens → Meaning



Lexical Semantics

- Lemma
- Wordforms
- Synonyms
- Antonyms
- Connotations
- Similar Tokens (Word Similarity)
- Related Tokens (Word Relatedness)
- Distribution (co-occurrences)

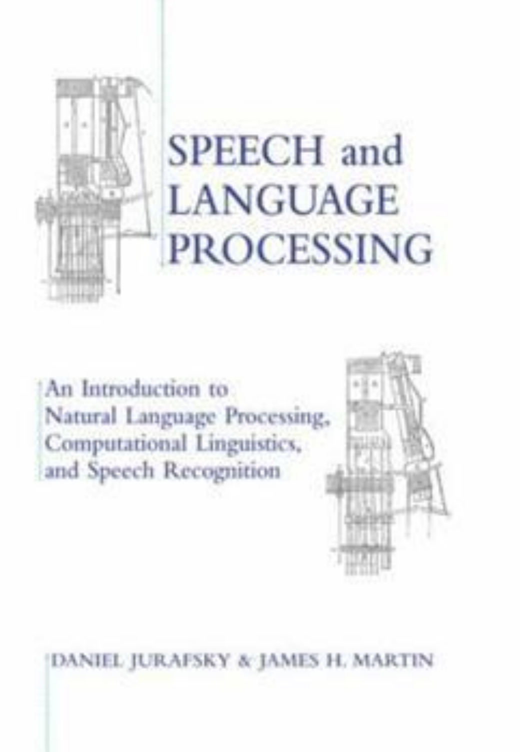
w : 'cat' ⇒



⇒ KG

No Vector Representation

Not very helpful to machines



Vector Semantics & Embeddings

CH06

Distributional Hypothesis

w ⇒ Doc ⇒ w w w

Words that occur in similar contexts tend to have similar meanings.
Meaning difference corresponds to difference in environments.

Joos, M. (1950). Description of language design. JASA, 22, 701–708.
Harris, Z. S. (1954). Distributional structure. Word, 10, 146–162.
Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955.

Vector Semantics

- Distribution (co-occurrences)
  - Count-based (Sparse Vector Representations)
    - Term-Doc Matrix
      - Tokens in Document Space
      - Documents in Token Space
    - Term-Term Matrix
      - Tokens in Token Space (co-occurrence context)
      - Document in Average of Constituent Tokens
    - Weighted Term-Term Matrix
      - Tf-Idf
  - Learning-based (Dense Vector Representations)
    - Matrix Factorization (Decomposition)
    - Classification (Word2Vec)

Doc

word → tf

row → dim

col →

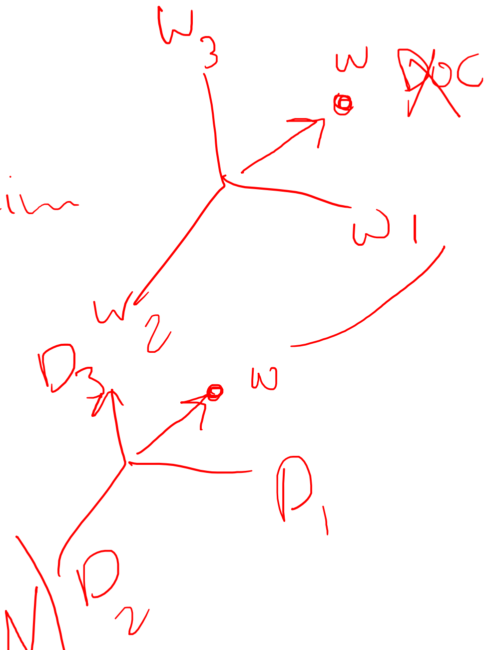
In word

w

w

DF

$\log\left(\frac{1}{DF} \times N\right)$



size dims

Applications

- Document Classification
  - Input vector to LR → sum → rand →
- User Classification?
- Movie Classification?
- Music Classification?

user → UGC

Imge

text

Audio

U_1

tweet

ϑ_{u_1}

Hoss

John

U_2

U_3

ϑ_{u_2}

Play Scr

imdb, wiki

text

movie →

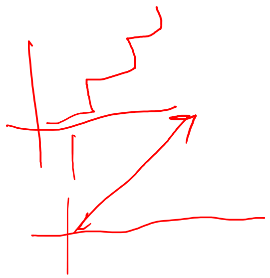
$m_1 \Rightarrow D_{m_1} \rightarrow \vartheta_{D_{m_1}}$

Vector Semantic Similarity/Distance

- Similarity = $\overline{\text{Distance}}$
  - Cosine Similarity
  - Minkowski Distance

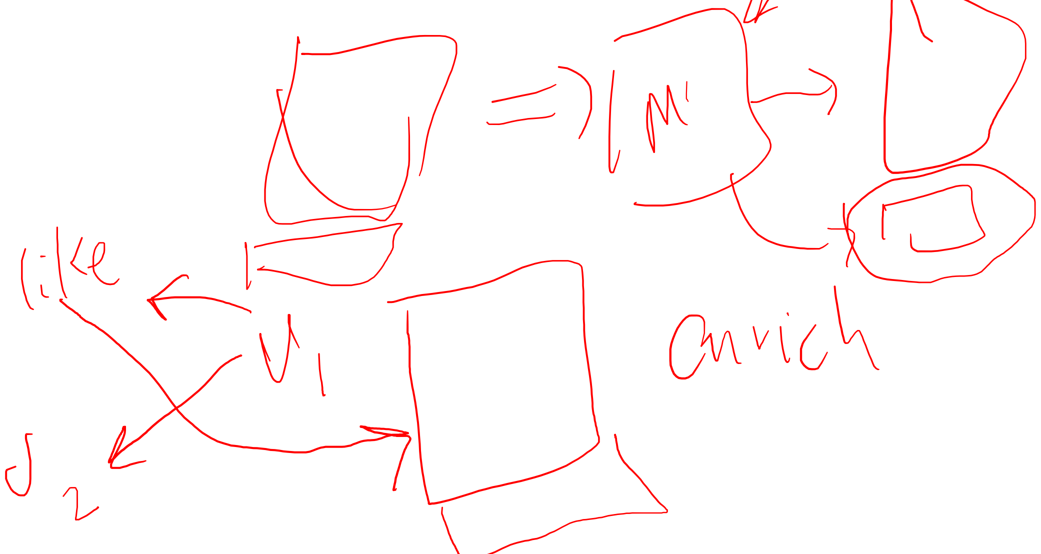
man

Eucl



Applications

- Document Clustering
- User Clustering?
- Information Retrieval
    - Query-based:
        - Search Engines (Document Recommendation)
    - 0-query-based (RecSys):
        - Friend Recommendation?
        - Movie Recommendation?
        - Music Recommendation?



Word2Vec

Learning Word Representations

Word2Vec as LR for w_i

Extreme Distributional Semantics: Bigrams

w_i w_{i+1} are semantically similar → Same Class (+)

w_i w_{i+2} are not semantically similar → Different Class (-)

$P(+| w_i w_{i+1}) = \text{sigmoid}[Vw_i:Vw_{i+1}][\text{Weights}] = 1.0$

$P(-|w_i w_{i+2}) = 1\text{-}P(+|w_i w_{i+2}) = 1 - \text{sigmoid}[Vw_i:Vw_{i+2}][\text{Weights}] = 1.0$

if we fix Vw_i → [Weights] of Vw_i

Word2Vec as LR for w_j

Extreme Distributional Semantics: Bigrams

w_j w_{j+1} are semantically similar → Same Class (+)

w_j w_{j+2} are not semantically similar → Different Class (-)

$P(+|\ w_j w_{j+1}) = \text{sigmoid}[Vw_j{:}Vw_{j+1}][\text{Weights}] = 1.0$

$P(-|w_j w_{j+2}) = 1\text{-}P(+|w_j w_{j+2}) = 1 - \text{sigmoid}[Vw_j{:}Vw_{j+2}][\text{Weights}] = 1.0$

if we fix Vw_j → [Weights] of Vw_j

Word2Vec as LR

Count-based

w_i w_k

w_j w_k

$$Vw_i \propto Vw_j$$

$$[\text{Weights}] \text{ of } Vw_i \propto [\text{Weights}] \text{ of } Vw_j$$

What can we tell about the [Weights]?

Word2Vec as LR

$$\mathrm{V}w_j \propto \mathrm{V}w_j$$

$$[\text{Weights}] \text{ of } \mathrm{V}w_i \;\cdot\; [\text{Weights}] \text{ of } \mathrm{V}w_j$$

What can we tell about the cosine of [Weights]?

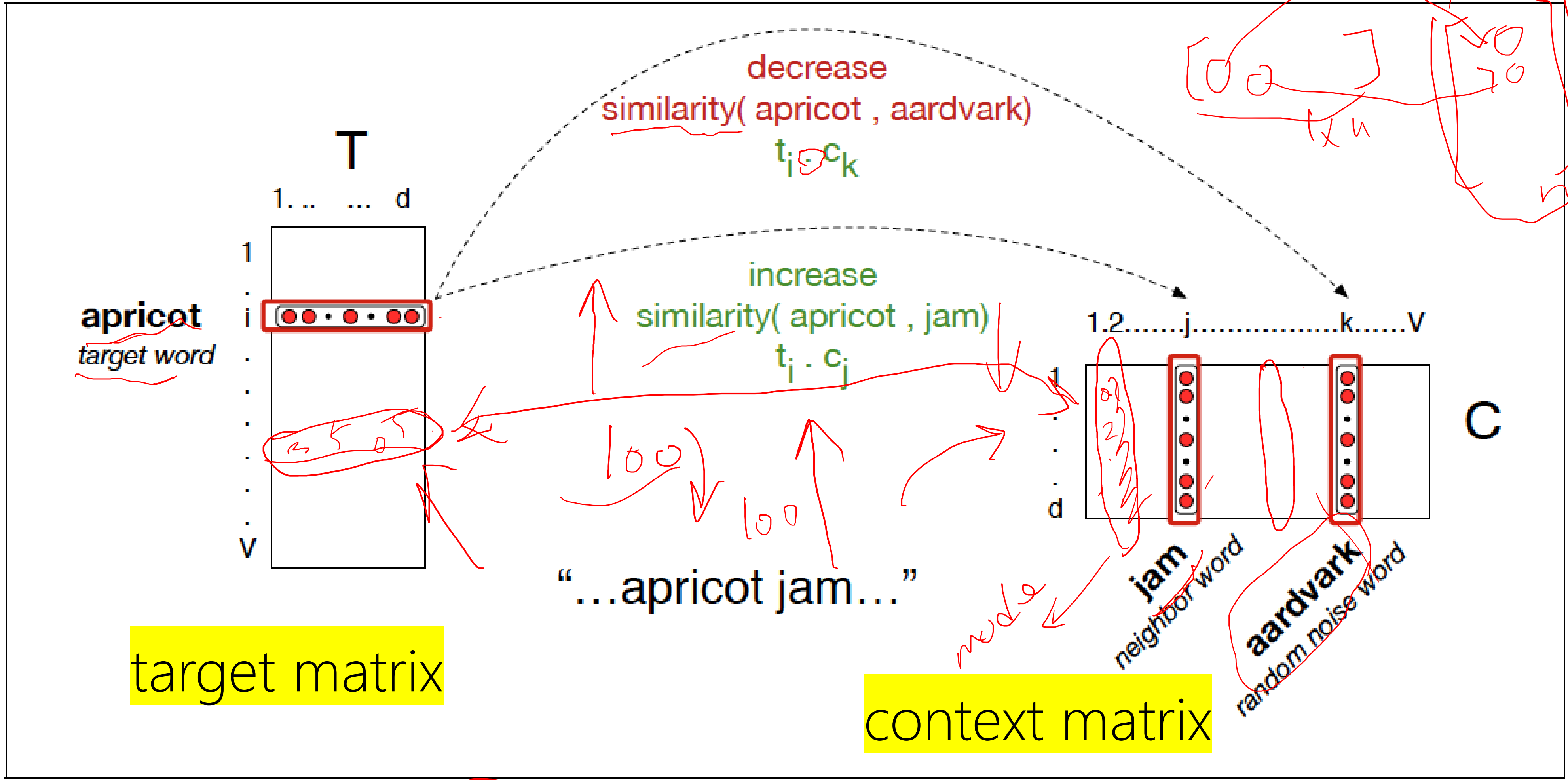


Figure 6.12 The skip-gram model tries to shift embeddings so the target embeddings (here for *apricot*) are closer to (have a higher dot product with) context embeddings for nearby words (here *jam*) and further from (have a lower dot product with) context embeddings for words that don't occur nearby (here *aardvark*).

Is it possible to use only one matrix?

Word2Vec

Given a context: … [tablespoon of apricot jam, a] …

Choose a word as target word *t:* apricot

Choose others as context word $c_{i:}$ jam, tablespoon

Estimate d-dimensional vectors for *t* and all c_i

Such that they are close to each other in d-dimensional space

d ≪ |Vocabs| or |Documents|

Close V_t *and* V_{c_i} → $V_t \cdot V_{c_i} > 0$ → $\sigma(V_t \cdot V_{c_i}) = \dfrac{1}{1+e^{-(V_w \cdot V_{c_i})}}$

Word2Vec

Given a context: … [tablespoon of apricot jam, a] …

Choose a word as target word *t:* apricot

Choose *random* word n_i from out of context: car, phone, …

apple car

Estimate d-dimensional vectors for *t* and all n_i

Such that they are far from each other in d-dimensional space

d ≪ |Vocabs| or |Documents|

car
Apple -Apple

distant V_t *and* V_{n_i} → $V_t \cdot V_{n_i} < 0$

$$V_t \cdot -V_{n_i} > 0 \rightarrow \sigma(V_t \cdot -V_{n_i}) = \frac{1}{1+e^{-(V_t \cdot -V_{n_i})}}$$

Word2Vec

Independent Assumption: P(x,y) = p(x)p(y)

$$P(+|t,c) = \frac{1}{1+e^{-t\cdot c}}$$

$$P(+|t,c_{1:k}) = \prod_{i=1}^{k} \frac{1}{1+e^{-t\cdot c_i}}$$

$$\log P(+|t,c_{1:k}) = \sum_{i=1}^{k} \log \frac{1}{1+e^{-t\cdot c_i}}$$

$$P(-|t,n_i) = \frac{1}{1+e^{-(t\cdot -ni)}}$$

$$P(-|t,n_{1:k}) = \prod_{i=1}^{k} \frac{1}{1+e^{-(t\cdot -ni)}}$$

$$\log P(-|t,n_{1:k}) = \sum_{i=1}^{k} \log \frac{1}{1+e^{-(t\cdot -ni)}}$$

$$L(\theta) = -\left(\sum_{(t,c)\in +} \log P(+|t,c) + \sum_{(t,c)\in -} \log P(-|t,c) \right)$$

$$= \frac{\partial L}{\partial W_i} \Rightarrow f(W_i)$$

Parameters

2d

Element of Matrices

Vectors of Words

v_{aT} v_{aC}

How many?

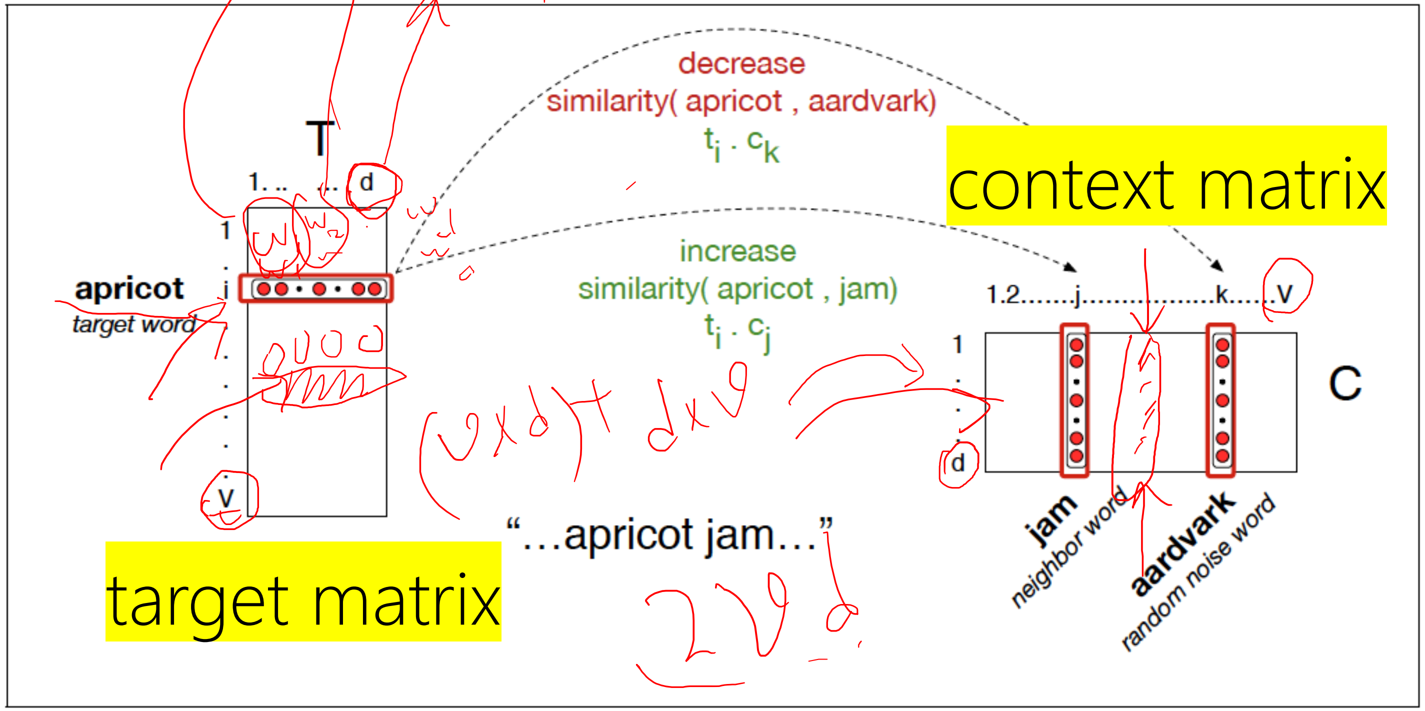


Figure 6.12 The skip-gram model tries to shift embeddings so the target embeddings (here for *apricot*) are closer to (have a higher dot product with) context embeddings for nearby words (here *jam*) and further from (have a lower dot product with) context embeddings for words that don't occur nearby (here *aardvark*).

Google Scholar

Tomas Mikolov

Senior Researcher, CIIRC CTU
Verified email at fb.com

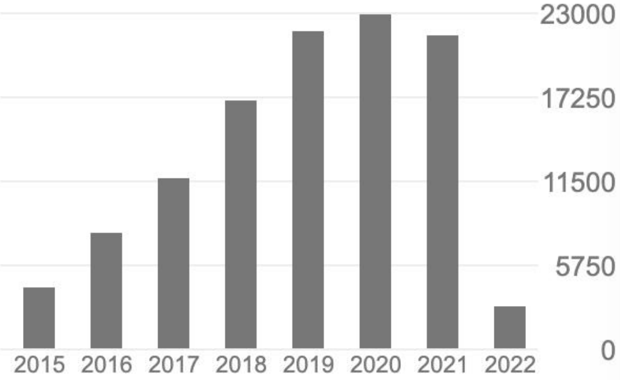
Artificial Intelligence   Machine Learning   Language Modeling   Natural Language Processing

FOLLOW

TITLE	CITED BY	YEAR
Distributed representations of words and phrases and their compositionality T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean Neural information processing systems	32433	2013
Efficient estimation of word representations in vector space T Mikolov, K Chen, G Corrado, J Dean arXiv preprint arXiv:1301.3781	27631	2013
Distributed representations of sentences and documents Q Le, T Mikolov International conference on machine learning, 1188-1196	9054	2014

Cited by   VIEW ALL

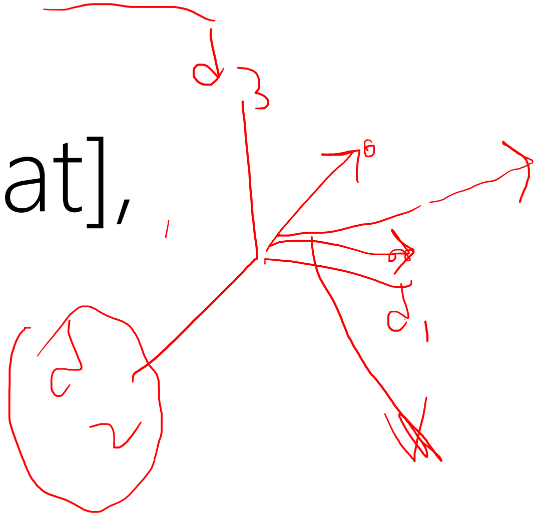
	All	Since 2017
Citations	113645	98036
h-index	46	44
i10-index	82	72



Word2Vec

- **Context Window?** Longer vs. Shorter?
- **Deterministic?** Any runs of training ended with same set of vectors?
- **Transformation?** rotation, flips, shear (skew), ...
- **Which signifier:**
    1. [cat], [miu], [*image_of_cat*], [ascii_cat],
    2. Count-based: [*tf*], [*tf-idf*], ...
    3. Learning methods: [word2vec]

(0.2, 0.5, -1.0)



Word2Vec → d, window

- **Problems?**

OOV

Set → hypermater

2

2000

user recom

music

IR

D_{u_i}

$u_i \Rightarrow$ P_1 P_2 P_3 … P_m $\Rightarrow \vartheta D_{u_i}$

D_{u_i} w_1 w_2 … w_4

$\vartheta_{w \in C}$ $\vartheta_{w \in C}$ $w \in C$

$\frac{1}{N} \sum_{w \in D} \vartheta_w \Rightarrow$ d

vector N

$N \Rightarrow$

dN

Fasttext (https://fasttext.cc)

- Word Vector Learning on Subwords
  - Char-level context
    - <natural>
      - c=~~2~~ 1 :<na, at, tu, ur, ra, al>
    - $V_{natural} = AVG(V_{bigram\text{-}chars}) = \frac{1}{6}(V_{na} + V_{at} + \cdots)$

v_a v_b v_c =>

apple =>

Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov; Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics 2017; 5 135–146. doi: https://doi.org/10.1162/tacl_a_00051

$v_w \leftrightarrow v_c$

Word2Vec as MF

If $Vw_i \propto Vw_j$ are given (e.g., tf-idf) → |V| |V| d

[Weights] of Vw_i ▪ [Weights] of Vw_j

$$|Vw_i \cdot Vw_j - ([\text{Weights}]_{Vwi} \cdot [\text{Weights}]_{Vwj})| \approx 0$$

10 10

Global Vectors (GloVe)

- Local Context + Global Context
  - Positive Point-wise Mutual Information (PPMI)

$$\hat{J} = \sum_{i,j} f(X_{ij})\left(w_i^T \tilde{w}_j - \log X_{ij}\right)^2$$

Read at Home

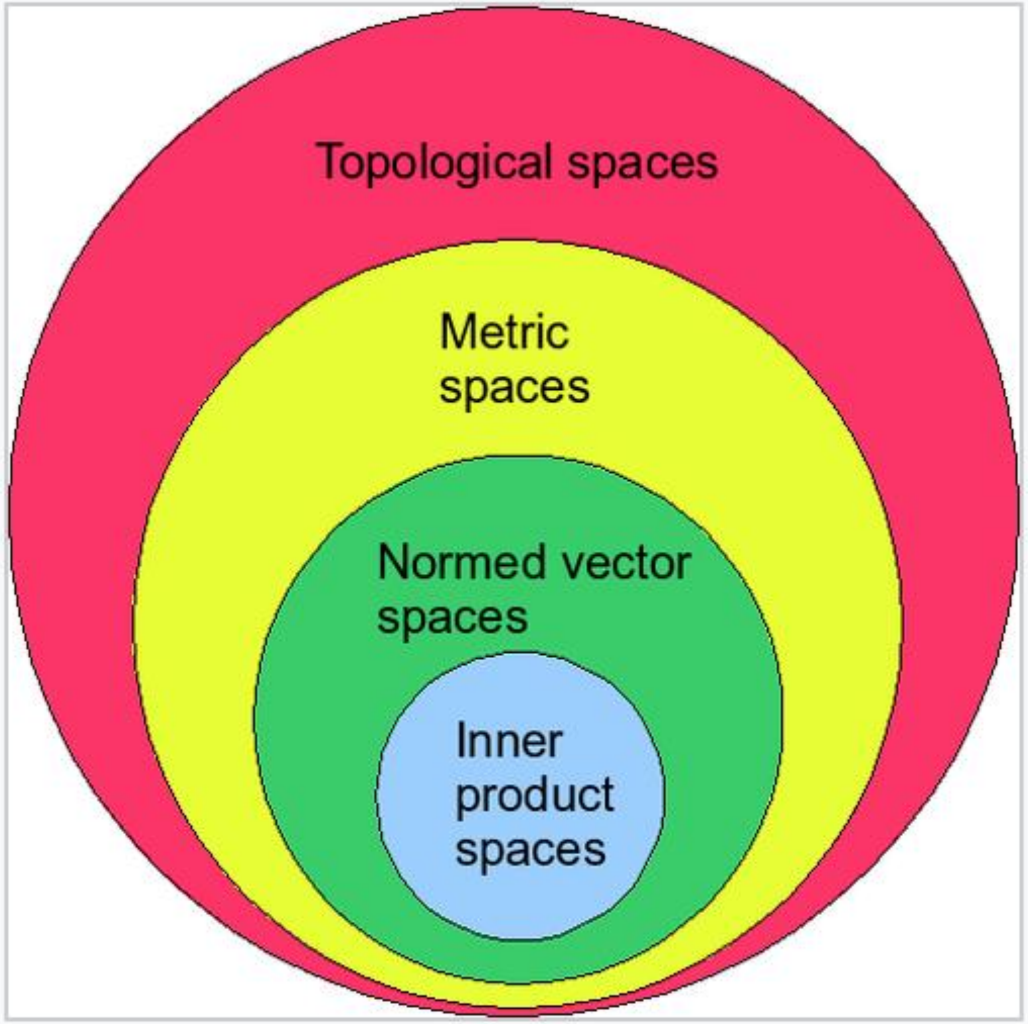
Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

Pre-trained Word Vectors

Available in genism python library:

- conceptnet-numberbatch-17-06-300 (1917247 records): ConceptNet Numberbatch consists of state...
- fasttext-wiki-news-subwords-300 (999999 records): 1 million word vectors trained on Wikipe...
- glove-twitter-100 (1193514 records): Pre-trained vectors based on 2B tweets,...
- glove-twitter-200 (1193514 records): Pre-trained vectors based on 2B tweets, ...
- glove-twitter-25 (1193514 records): Pre-trained vectors based on 2B tweets, ...
- glove-twitter-50 (1193514 records): Pre-trained vectors based on 2B tweets, ...
- glove-wiki-gigaword-100 (400000 records): Pre-trained vectors based on Wikipedia 2...
- glove-wiki-gigaword-200 (400000 records): Pre-trained vectors based on Wikipedia 2...
- glove-wiki-gigaword-300 (400000 records): Pre-trained vectors based on Wikipedia 2...
- glove-wiki-gigaword-50 (400000 records): Pre-trained vectors based on Wikipedia 2...
- word2vec-google-news-300 (3000000 records): Pre-trained vectors trained on a part of...
- word2vec-ruscorpora-300 (184973 records): Word2vec Continuous Skipgram vectors tra...

Vector Semantics
Vector Space
Transformation
Linear Algebra

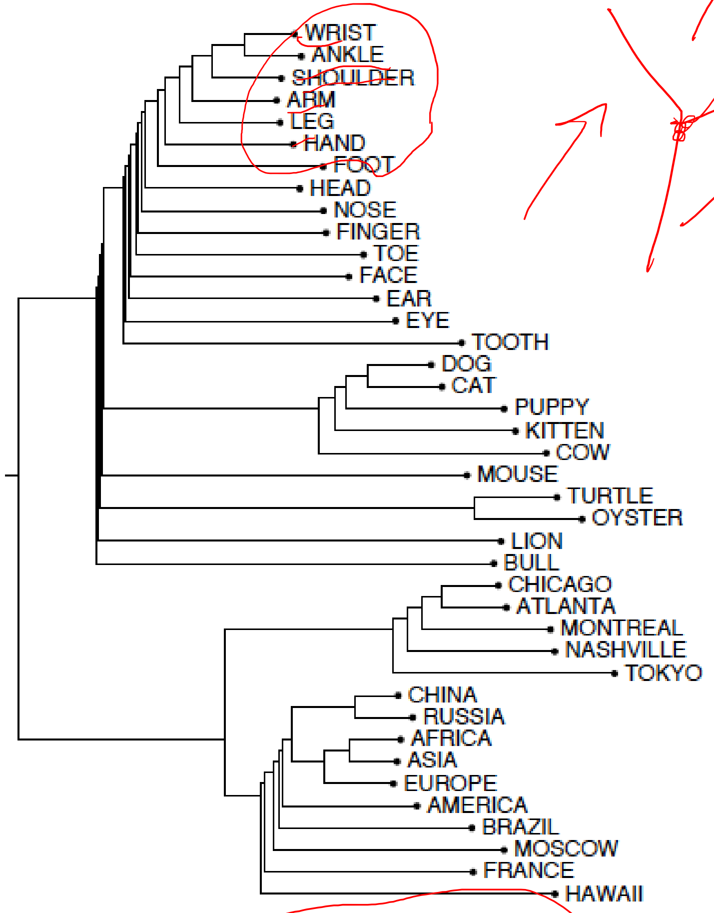


Hierarchy of mathematical spaces. Normed vector spaces are a superset of inner product spaces and a subset of metric spaces, which in turn is a subset of topological vector space.

https://www.youtube.com/watch?v=Dmc3mQ87GiQ

Linguistic ↔ Geometric

(man - wom = king - queen   d=100 ⇒ tsne → 2-d
3-d



Hierarchy

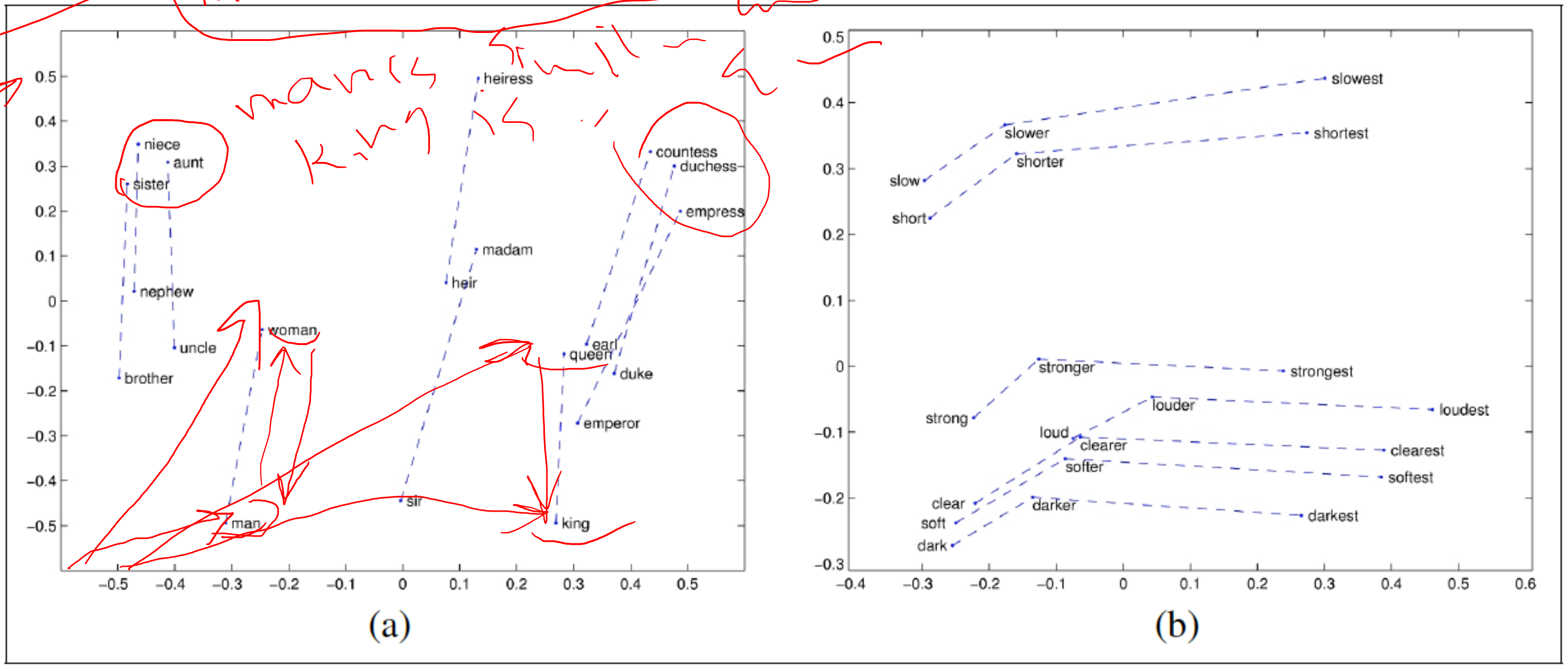


Figure 6.13 Relational properties of the vector space, shown by projecting vectors onto two dimensions. (a) 'king' - 'man' + 'woman' is close to 'queen' (b) offsets seem to capture comparative and superlative morphology (Pennington et al., 2014).

Linguistic ↔ Geometric

Analogy

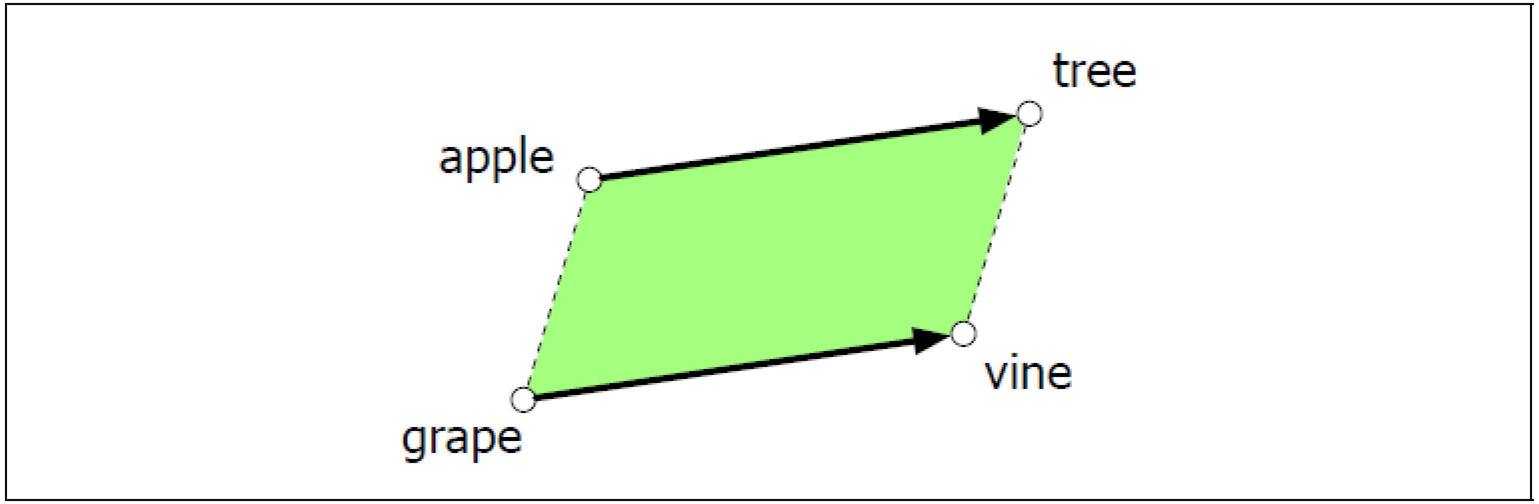


Figure 6.15 The parallelogram model for analogy problems (Rumelhart and Abrahamson, 1973): the location of $\overrightarrow{\text{vine}}$ can be found by subtracting $\overrightarrow{\text{apple}}$ from $\overrightarrow{\text{tree}}$ and adding $\overrightarrow{\text{grape}}$.

a to *b* is the same as c to ?

Movement

Temporality (How?)

tim 1900

2000

W-1

W-2000

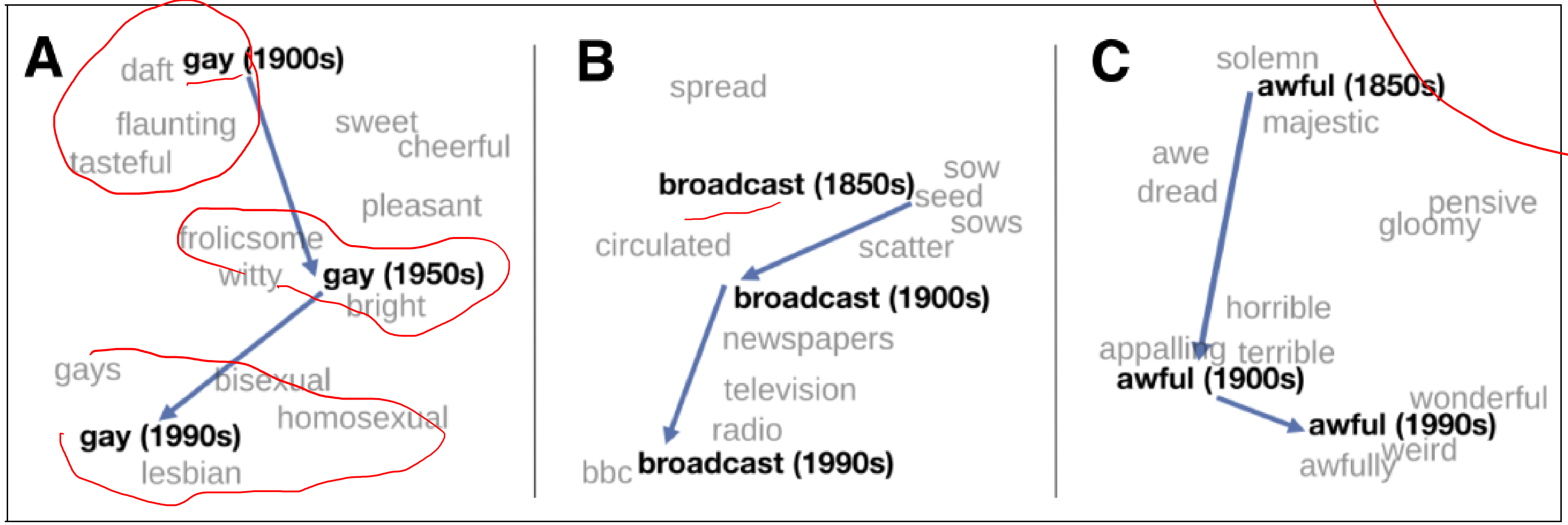


Figure 6.14 A t-SNE visualization of the semantic change of 3 words in English using word2vec vectors. The modern sense of each word, and the grey context words, are computed from the most recent (modern) time-point embedding space. Earlier points are computed from earlier historical embedding spaces. The visualizations show the changes in the word *gay* from meanings related to "cheerful" or "frolicsome" to referring to homosexuality, the development of the modern "transmission" sense of *broadcast* from its original sense of sowing seeds, and the pejoration of the word *awful* as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Hamilton et al., 2016b).

Biases

Product Re

Inherent/Latent/Hidden Distribution

- (sare, mom, nurse), (mr., ahmed, doctor, president)
- (drug, mexican), (education, usa, canada)
- (flowers, pleasant, {European-American}), (insects, ugly, {African-American})

Debiasing

- Gender-base: [he] remains masculine, [she] remains feminine, but [nurse],[doctor],[president] becomes neutral

Study of Bias in History

Evaluation

lexical semantic ⇐ Apple — orange ⇒ v_{apple} v_{orange}

Intrinsic

- Golden Standards for Semantic Similarity/Distance
  - No Context: just pair of words
    - WordSim-353
    - SimLex-999
  - With Context:
    - Stanford Contextual Word Similarity (SCWS) (Huang et al., 2012) and the
    - Word-in-Context (WiC) (Pilehvar and Camacho-Collados, 2019)

top K - Apple: orange, corn

top K - apple: orange, tree

Apple corn

Extrinsic:

- Improve the performance of underlying task
  - Information Retrieval (IR), Document Classification, Sentiment Analysis, ...

CROSS-LINGUAL WORD EMBEDDINGS

Words from two or more languages are represented in the same shared low-dimensional vector space.
Level of supervision:

- sentence-level: Machine Translation (MT) Corpora
- document-level: Wikipedia
- word-level: Bilingual Dictionaries
- unsupervised: Distribution of words in monolingual corpora in a bilingual dictionary

a p o

How to learn representation for:

- Character (autocorrection)
- Sentence/Paragraph/Documents (Doc2Vec)
  - AVG is predefined function over the constituent words
  - Let's learn the aggregation function from data!

awesome

User2Vec => *2Vec

music2Vec

image2Ve