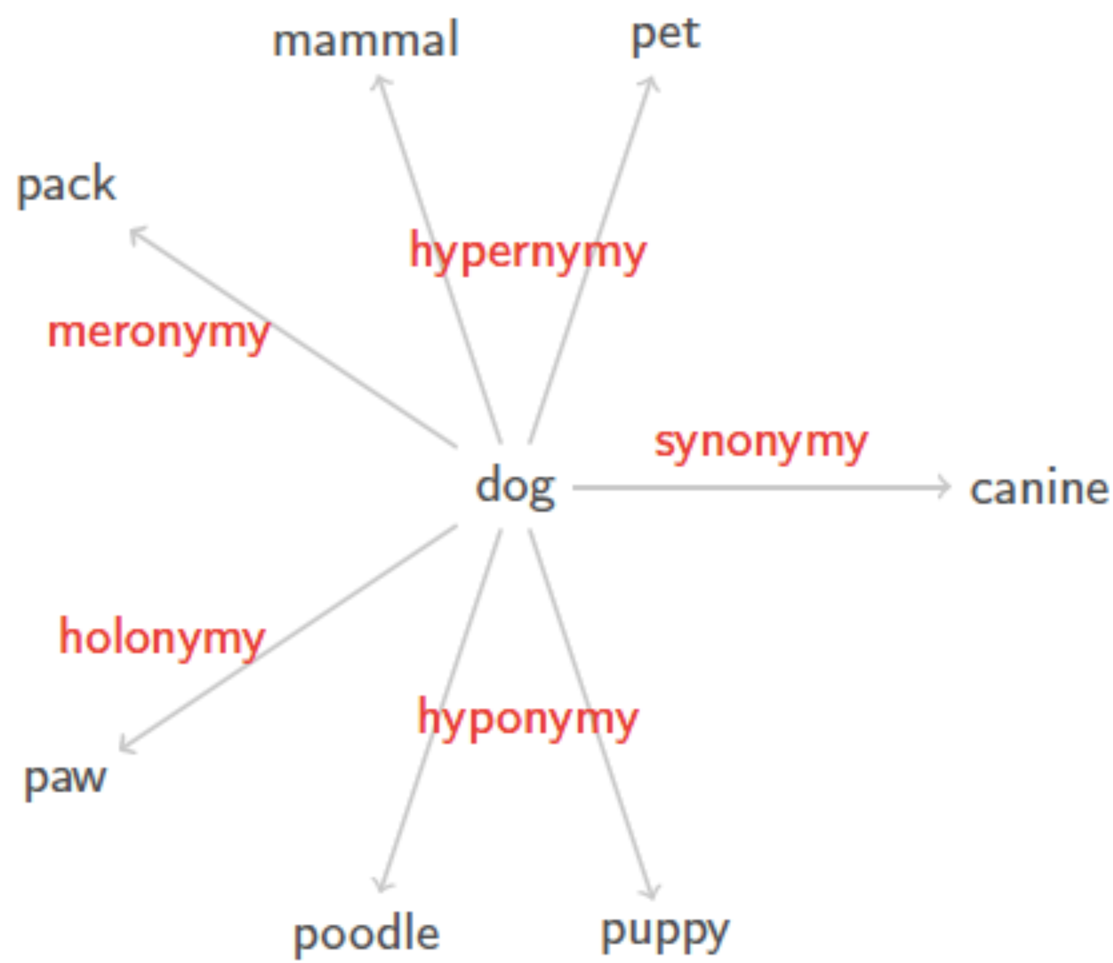
Distributional Semantics and Word Embeddings

Announcements

- Midterm returned at end of class today
  - Only exams that were taken on Thursday
- Today: moving into neural nets via word embeddings
- Tuesday: Introduction to basic neural net architecture. Chris Kedzie to lecture.
- Homework out on Tuesday
- Language applications using different architectures

Lexical semantics

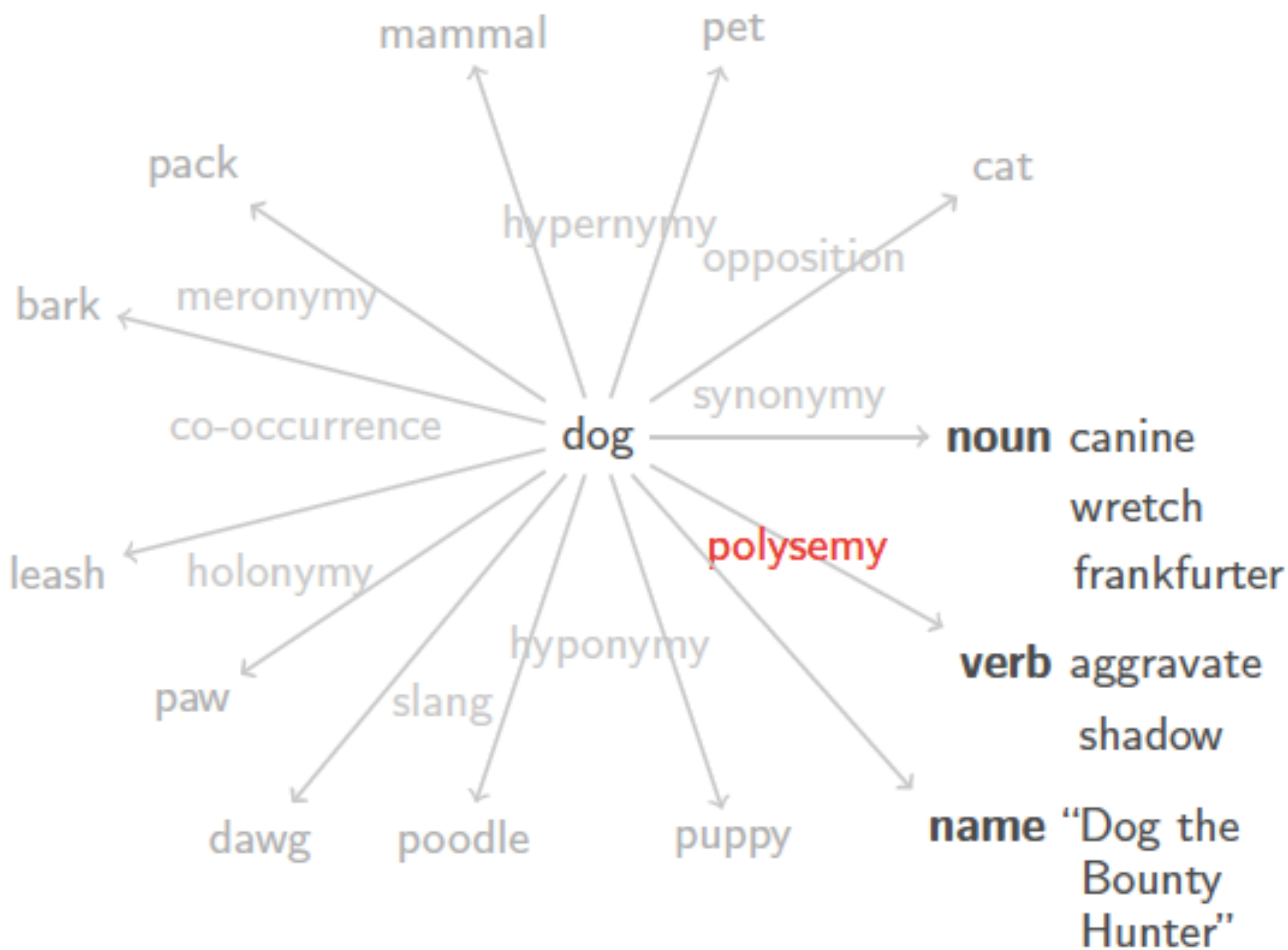


Slide from Kapil Thadani

Lexical semantics

Slide from Kapil Thadani

Lexical semantics



Slide from Kapil Thadani

Methods so far

- WordNet: an amazing resource.. *But*

- *What are some of the disadvantages?*

Methods so far

- Bag of words
  - Simple and interpretable

- In vector space, represent a sentence

  John likes milk

  [0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0]

  "one-hot" vector

  Values could be frequency, TF*IDF

- Sparse representation
  - Dimensionality: 50K unigrams, 500K bigrams

- Curse of dimensionality!

From Symbolic to Distributed Representations

- Its problem, e.g., for web search
  - If user searches for [Dell notebook battery], should match documents with “Dell laptop battery”
  - If user searches for [Seattle motel] should match documents containing “Seattle hotel”
- But
  - Motel [0 0 0 0 0 0 0 0 0 0 0 1 0 0 0]
    Hotel [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
- Our query and document vectors are orthogonal
  There is no natural notion of similarity in a set of one-hot vectors

- -> Explore a direct approach where vectors encode it

Slide from Chris Manning

Distributional Semantics

- "You shall know a word by the company it keeps" [J.R. Firth 1957]
  - *Marco saw a hairy little wampunuk hiding behind a tree*

- Words that occur in similar contexts have similar meaning

- Record word co-occurrence within a window over a large corpus

Word Context Matrices

- Each row_i represents a word
- Each $column_j$ represents a linguistic context
- $Matrix_{ij}$ represents strength of association
  - $M^f \varepsilon R$, $M^f_{i,j} = f(w_i, c_j)$ where f is an association measure of the strength between a word and a context

	I	hamburger	book	gift	spoon
ate	.45	.56	.02	.03	.3
gave	.46	.13	.67	.7	.25
took	.46	.1	.7	.5	.3

Associations and Similarity

- Effective association measure: Pointwise Mutual Information (PMI)

  $$\log\ P(w,c)/P(w)P(c)$$
  $$= \log\ \#(w,c)*|D|/\#(w)*\#(c)$$

- Compute similarity between words and text
  - Cosine Similarity

    $$\Sigma_i\, u_i \bullet v_i \,/\, \surd\Sigma_i(u_i)^2\ \surd\Sigma_i(v_i)^2$$

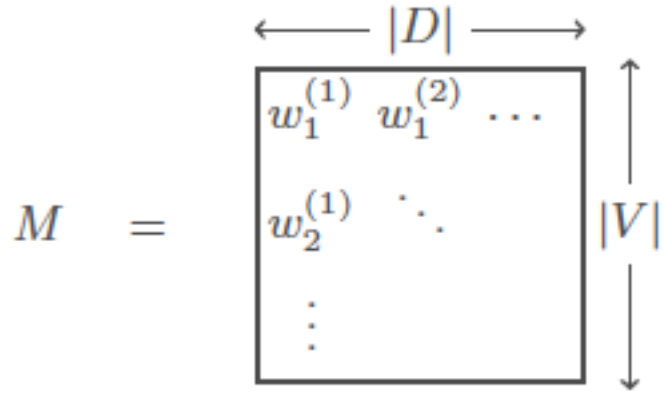
Dimensionality Reduction

- Captures context, but still has sparseness issues
- Singular value decomposition (SVD)
  - Factors matrix M into two narrow matrices: W, a word matrix, and C, a context matrix such that $WC^T = M'$ is the best rank-d approximation of M
- A “smoothed” version of M
  - Adds words to contexts if other words in this context seem to co-locate with each other
  - Represents each word as a dense d-dimensional vector instead of a sparse $|V_C|$ one

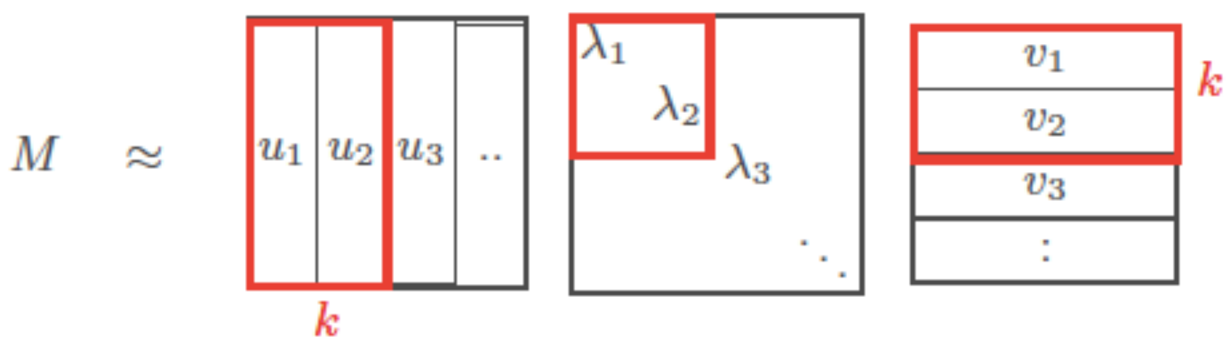
Latent Semantic Analysis

Deerwester et al. (1990)

Construct term-document matrix



Singular value decomposition



Select top k singular vectors for k-dim embeddings of words/docs

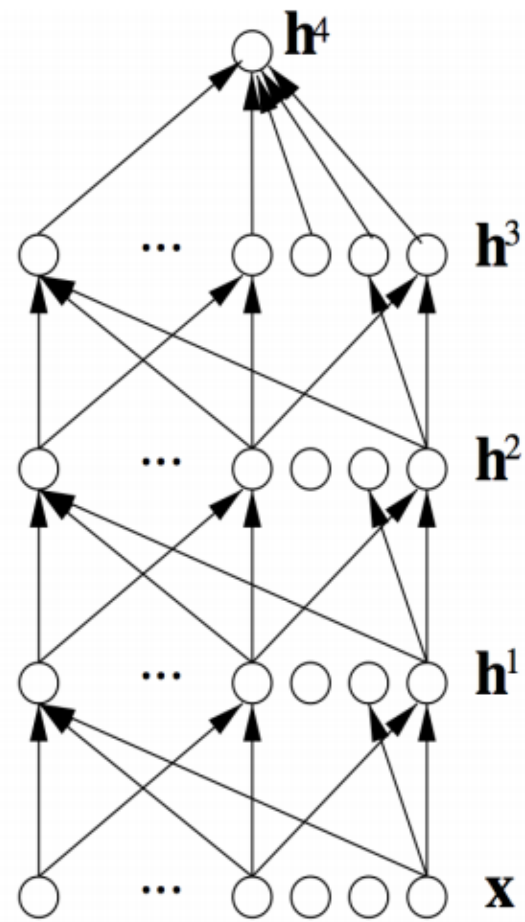
Slide from Kapil Thadani

Neural Nets

- A family of models within deep learning
- The machine learning approaches we have seen to date rely on “feature engineering”
- With neural nets, instead we learn by optimizing a set of parameters

Why "Deep Learning"?

- *Representation learning* attempts to automatically learn good features or representations
- *Deep learning* algorithms attempt to learn (multiple levels of) representation and an output
- From "raw" inputs x (e.g., sound, characters, words)



Slide adapted from Chris Manning

Reasons for Exploring Deep Learning

- Manually designed features can be over-specific or take a long time to design
  - … but can provide an intuition about the solution

- Learned features are easy to adapt

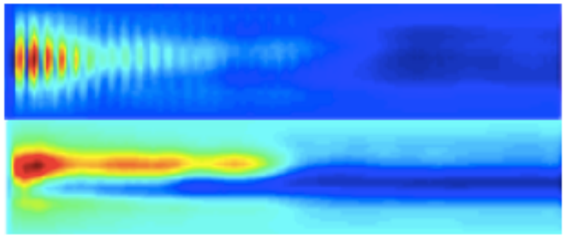
- Deep learning provides a very flexible framework for representing word, visual and linguistic information

- Both supervised and unsupervised methods

Slide adapted from Chris Manning

Progress with deep learning

- Huge leaps forward with
  - Speech



  - Vision

  - Machine Translation

[Krizhevsky et al. 2012]

  - More modest advances in other areas

From Distributional Semantics to Neural Networks

- Instead of count-based methods, distributed representations of word meaning

- Each word associated with a vector where meaning is captured in different dimensions as well as in dimensions of other words

- Dimensions in a distributed representation are not interpretable

- Specific dimensions do not correspond to specific concepts

Basic Idea of Learning Neural Network Embeddings

- Define a model that aims to predict between a center word w_t and context words in terms of word vectors

  $p(\text{context}|w_t)=....$

  Which has a loss function, e.g.,

  $J = 1- p(w_{-t|w}t)$

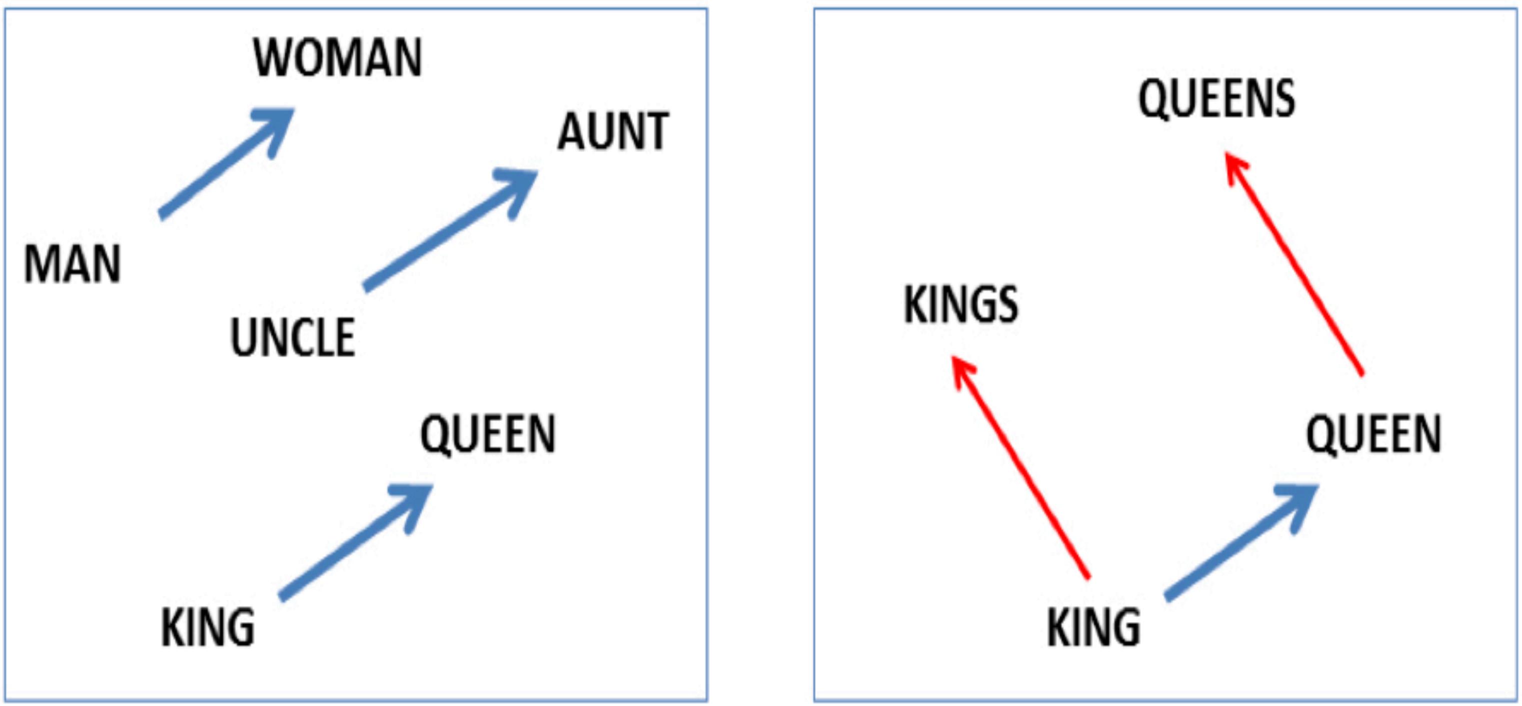
- We look at many positions t in a large corpus

- We keep adjusing the vector representations of words to minimize loss

Slide adapted from Chris Manning

Embeddings Are Magic

vector(*'king'*) - vector(*'man'*) + vector(*'woman'*) ≈ vector('queen')



Slide from Dragomir Radev, Image courtesy of Jurafsky & Martin

Relevant approaches: Yoav and Goldberg

- Chapter 9: A neural probabilistic language model (Bengio et al 2003)

- Chapter 10, p. 113 NLP (almost) from Scratch (Collobert & Weston 2008)

- Chapter 10, p 114 Word2vec (Mikolog et al 2013)

Main Idea of word2vec

- Predict between every word and its context

- Two algorithms
  - Skip-gram (SG)
    Predict context words given target (position independent)
  - Continuous Bag of Words (CBOW)
    Predict target word from bag-of-words context

Slide adapted from Chris Manning

Training Methods

- Two (moderately efficient) training methods

  Hierarchical softmax

  Negative sampling

  Today: naïve softmax

Slide adapted from Chris Manning

Instead, a **bank** can hold the investments in a custodial account

Context words | center word | context words
2 word window | t | 2 word window

But as agriculture burgeons on the east **bank**, the river will shrink

Context words | center | context
2 word window | t | 2 word window

Objective Function

- Maximize the probability of context words given the center word

$$J'(\Theta) = \prod_{t=1} \prod_{\substack{-m \le j \le m \\ j \ne 0}} P(w_{t+j} \mid w_{t\ j}\ \Theta)$$

Negative log likelihood

$$J'(\Theta) = -1/T \sum_{t=1} \sum_{\substack{-m \le j \le m \\ j \ne 0}} \log P(w_{t+j} \mid w_t)$$

Where Θ represents all variables to be optimized

Slide adapted from Chris Manning

Softmax

using word c to obtain probability of word o

- Convert $P(w_{t+j} | w_t)$

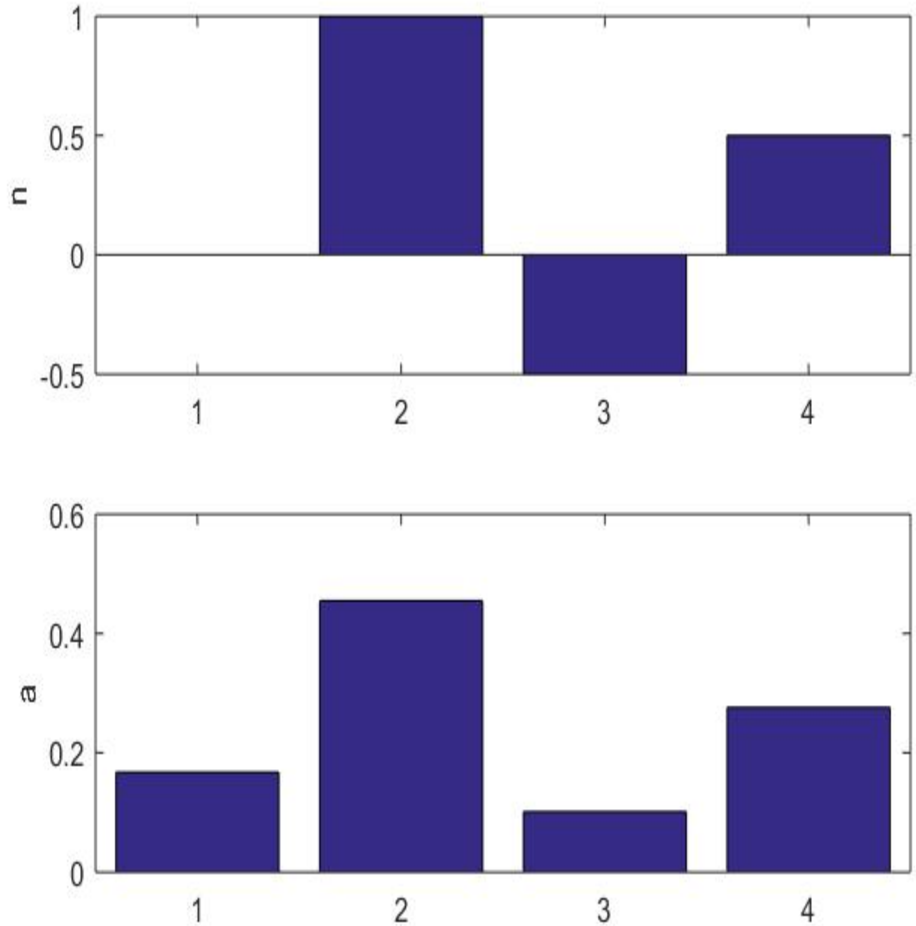
$$P(o|c) = \exp(u_o^T \, v_c) / \Sigma^{v}_{w=1} \exp(u_w^T v_c)$$

exponentiate to make positive   normalize

where o is the outside (or output) word index and c is the center word index, v_c and u_o are center and outside vectors of indices c and o

Slide adapted from Chris Manning

Softmax



Slide from Dragomir Radev

Dot Product

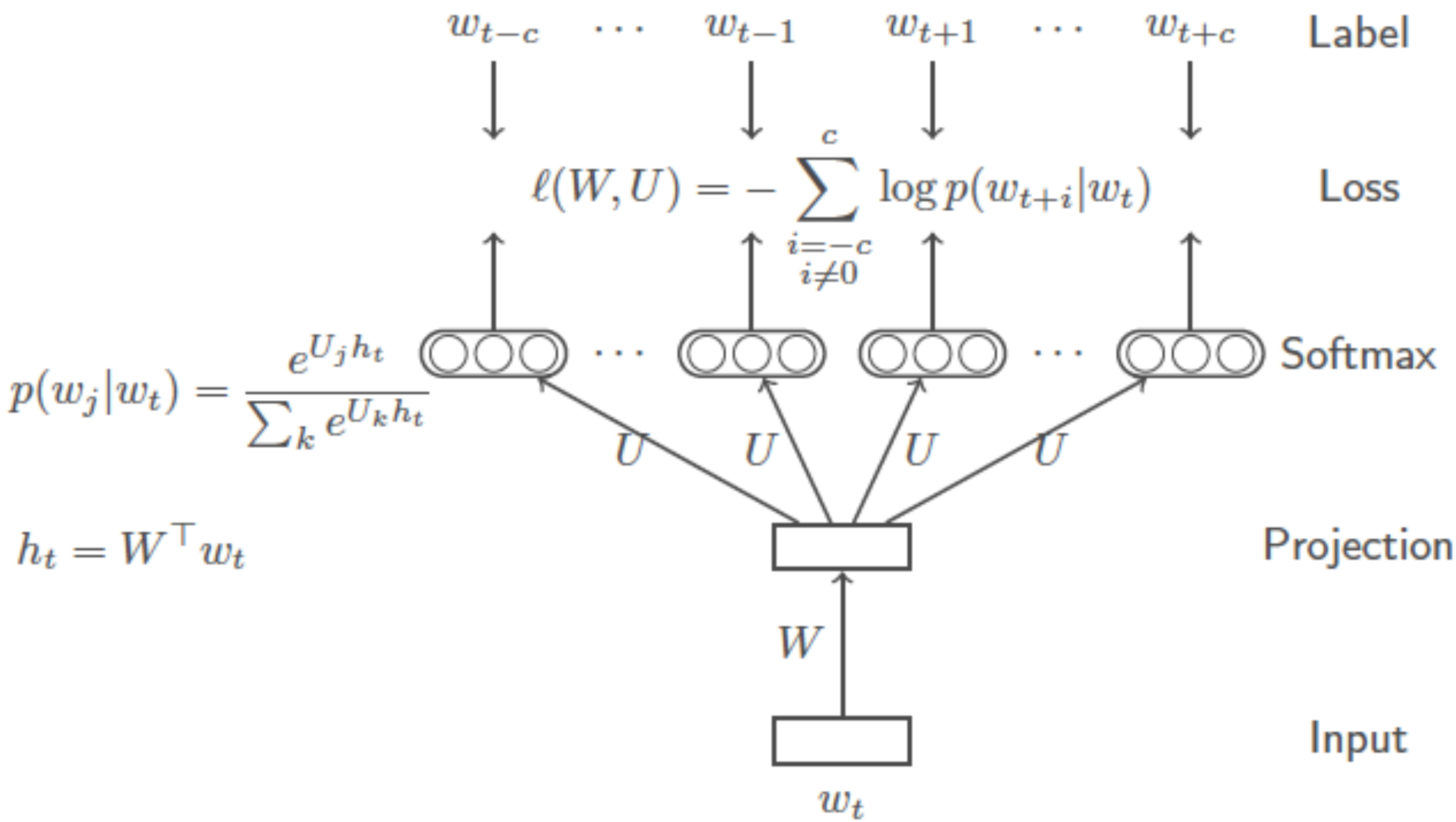
- $u^T v = u \bullet v = \Sigma^n_{i=1} u_i v_i$
- Bigger if u and v are more similar

word2vec

Mikolov et al. (2013)

Skip-gram

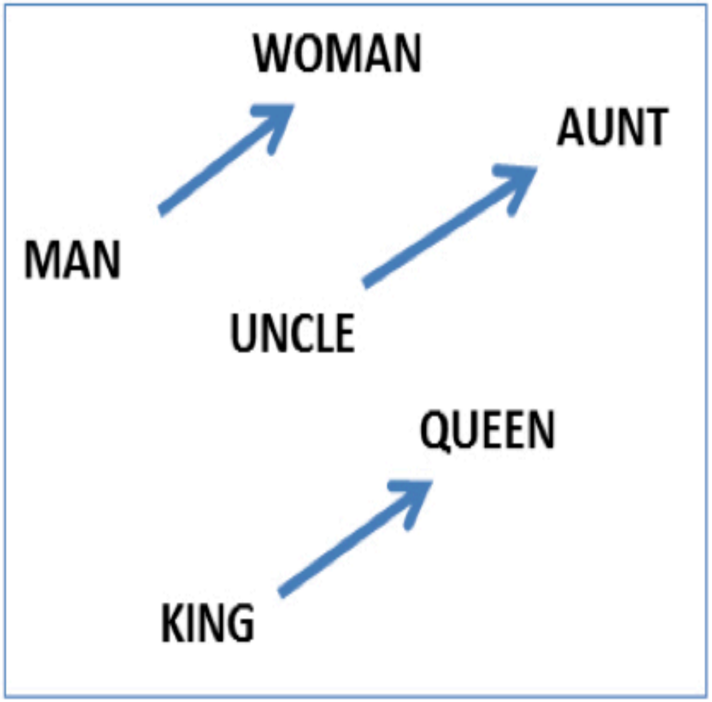
- Predict context $w_{t-c}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+c}$ given target w_t

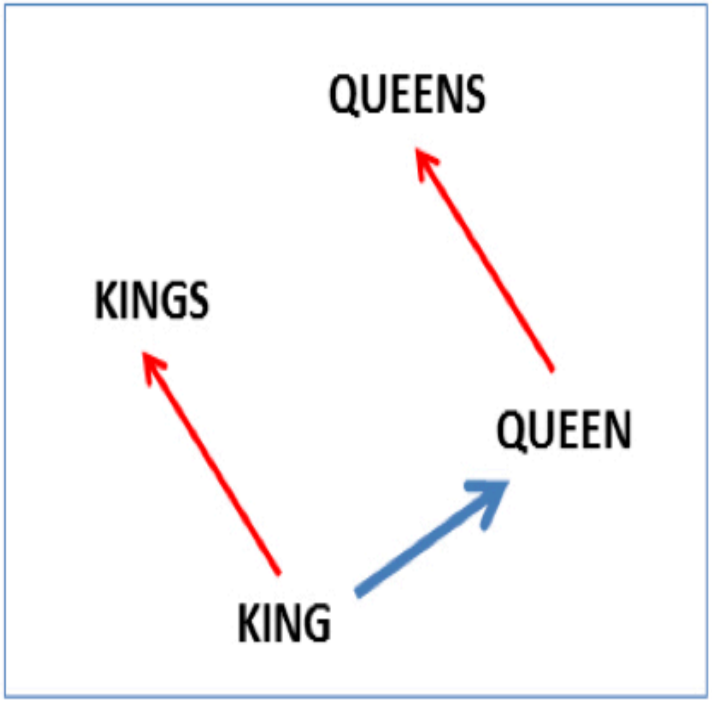


Slide from Kapil Thadani

Embeddings Are Magic

vector(*'king'*) - vector(*'man'*) + vector(*'woman'*) $\approx$ vector('queen')





Slide from Dragomir Radev, Image courtesy of Jurafsky & Martin

ord2vec

Mikolov et al. (2013)

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Additive compositionality

Evaluating Embeddings

- Nearest Neighbors
- Analogies
  - (A:B)::(C:?)
- Information Retrieval
- Semantic Hashing

Slide from Dragomir Radev

Similarity Data Sets

Dataset	Word pairs	Reference
RG	65	Rubenstein and Goodenough (1965)
MC	30	Miller and Charles (1991)
WS-353	353	Finkelstein et al. (2002)
YP-130	130	Yang and Powers (2006)
MTurk-287	287	Radinsky et al. (2011)
MTurk-771	771	Halawi et al. (2012)
MEN	3000	Bruni et al. (2012)
RW	2034	Luong et al. (2013)
Verb	144	Baker et al. (2014)
SimLex	999	Hill et al. (2014)

[Table from Faruqui et al. 2016]

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

[Mikolov et al. 2013]

Semantic Hashing

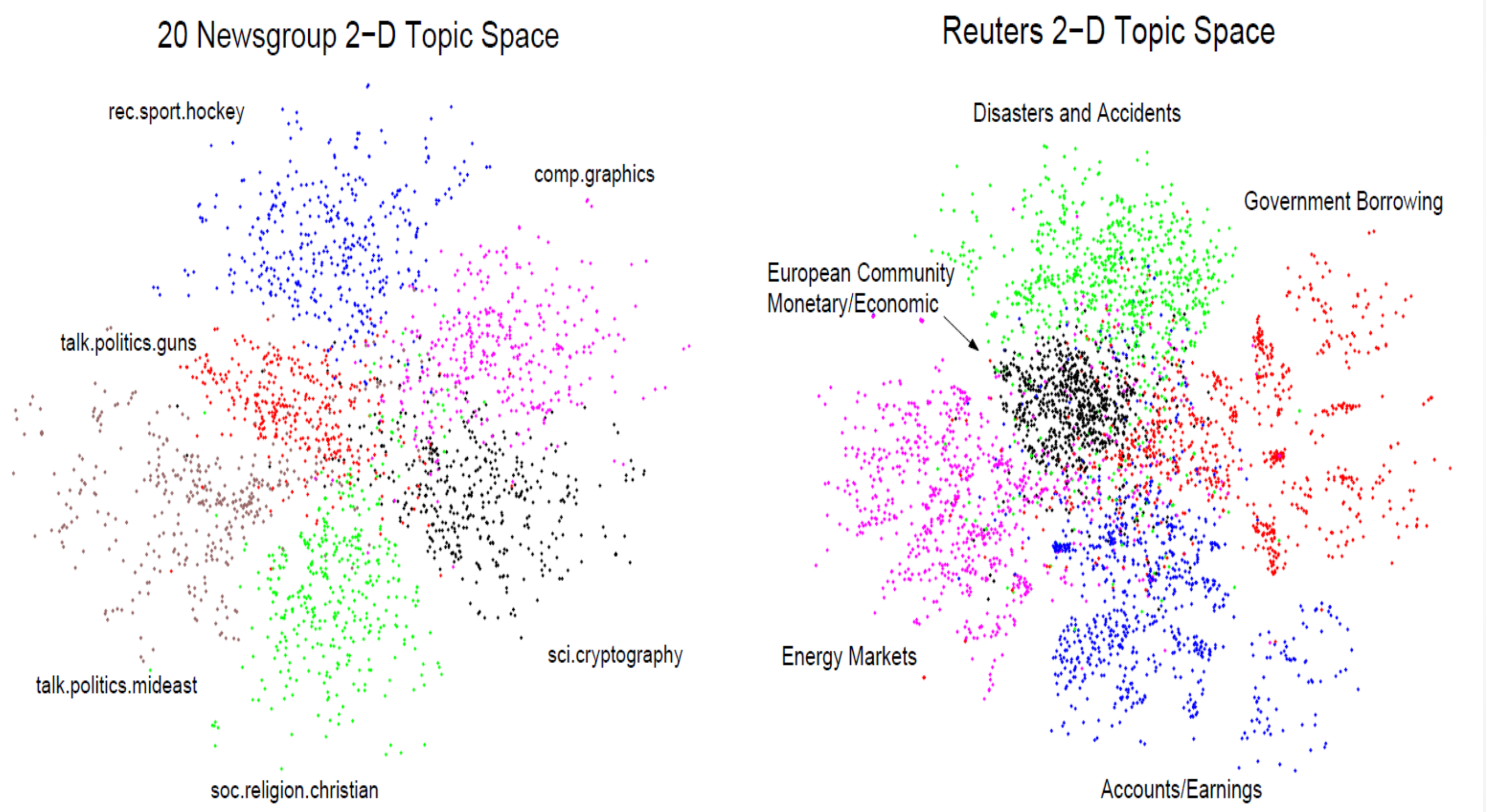
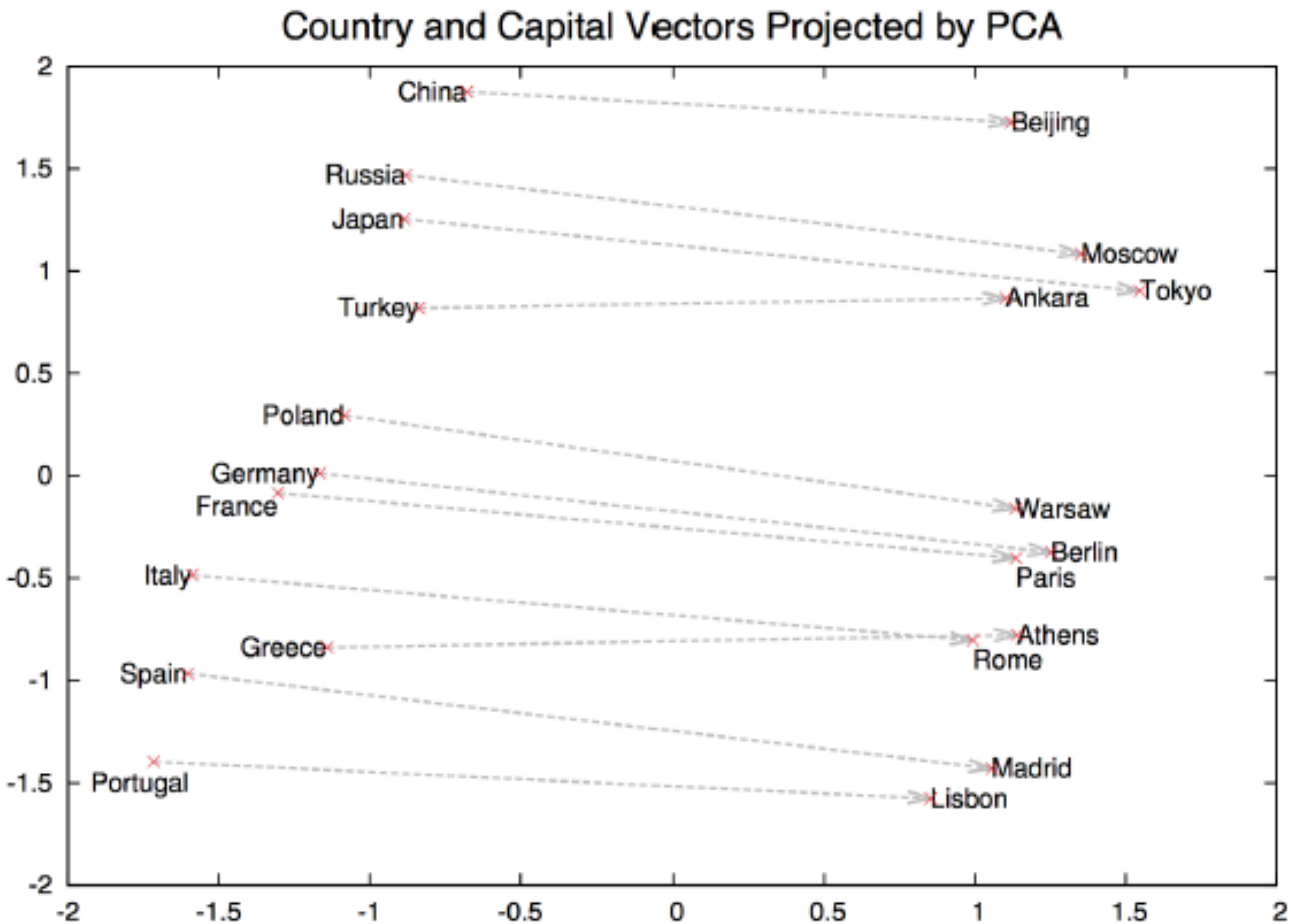


Figure 5: A 2-dimensional embedding of the 128-bit codes using stochastic neighbor embedding for the 20 Newsgroups data (left panel) and the Reuters RCV2 corpus (right panel). See in color for better visualization.

[Salakhutdinov and Hinton 2

word2vec

Mikolov et al. (2013)



Visualizing lexical relationships

word2vec

Mikolov et al. (2013)

Newspapers			
New York	New York Times	Baltimore	Baltimore Sun
San Jose	San Jose Mercury News	Cincinnati	Cincinnati Enquirer
NHL Teams			
Boston	Boston Bruins	Montreal	Montreal Canadiens
Phoenix	Phoenix Coyotes	Nashville	Nashville Predators
NBA Teams			
Detroit	Detroit Pistons	Toronto	Toronto Raptors
Oakland	Golden State Warriors	Memphis	Memphis Grizzlies
Airlines			
Austria	Austrian Airlines	Spain	Spainair
Belgium	Brussels Airlines	Greece	Aegean Airlines
Company executives			
Steve Ballmer	Microsoft	Larry Page	Google
Samuel J. Palmisano	IBM	Werner Vogels	Amazon

Phrase analogies

How are word embeddings used?

- As features in supervised systems

- As the main representation with a neural net application/task

Are Distributional Semantics and Word Embeddings all that different?

Homework2

- Max 99.6, Min 4, Stdev: 21.4
- Mean 82.2, Median 92.1
- Vast majority of F1 scores between 90 and 96.5.

Midterm

- Max: 95, Min: 22.5
- Mean: 66.6, Median 68.5
- Standard Deviation: 15
- Will be curved and the curve will be provided in the next lecture