

Word Embeddings

NLP
Andreas Marfurt

Overview

- Distributional hypothesis
- Low-dimensional word embedding methods
 - Word2vec
 - GloVe
 - FastText
- Use cases
 - Word analogy
 - Downstream applications

Distributional Hypothesis

"You shall know a word by the company it keeps."
— Firth, J.R. (1957)

(Also attributed to [Harris, 1954](#))

Distributional Hypothesis

"You shall know a word by the company it keeps."

- Words are characterized by their context
- A word appears often in the same context
 - We can deduce its meaning from the context
 - If a word has multiple meanings, we expect that the contexts differ, too (polysemy)

The dogs played on the _____ in the park.

Distributional Hypothesis

"You shall know a word by the company it keeps."

- Words are characterized by their context
- A word appears often in the same context
 - We can deduce its meaning from the context
 - If a word has multiple meanings, we expect that the contexts differ, too (polysemy)

The dogs played on the _____ in the park.

- grass
- lawn

Distributional Hypothesis

"You shall know a word by the company it keeps."

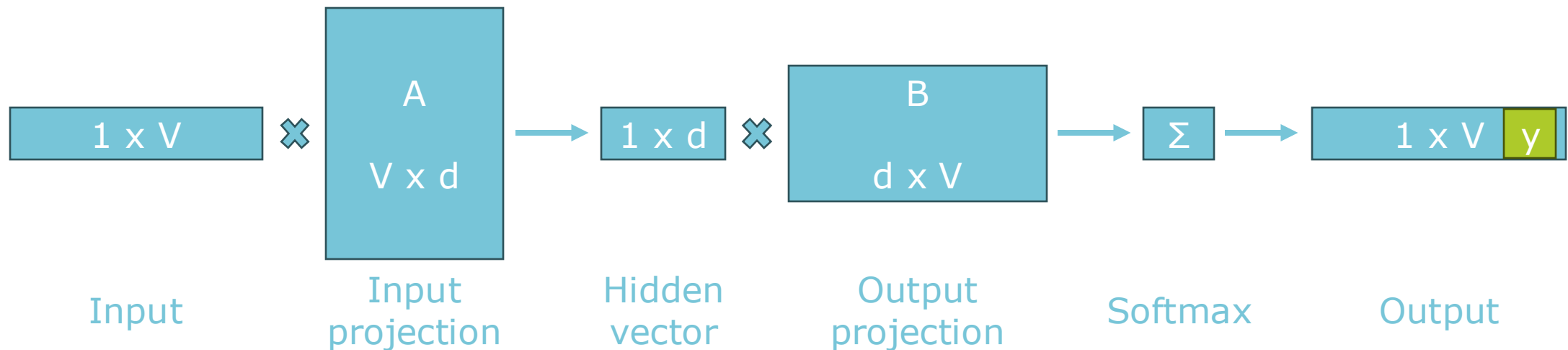
- Roger Federer won another _____ tournament.
- Now put the cake into the oven and let it _____ for 20 minutes.
- She was quite upset, _____ she decided it was better not to let it show.

Word2vec

- Mikolov et al., 2013 [[a](#), [b](#)]
- Idea: Compute word vectors (= representations/embeddings) for each word from their contexts
- Requirements
 - Want to use a large corpus, must be efficient
 - Don't have labels, must be unsupervised (or self-supervised)

Word2vec Architecture

- 2-layer neural network
- Input: one-hot word vector, V : vocabulary size
- A : Projection to inner dimension d
- B : Projection to output
- Softmax Σ to predict target word y

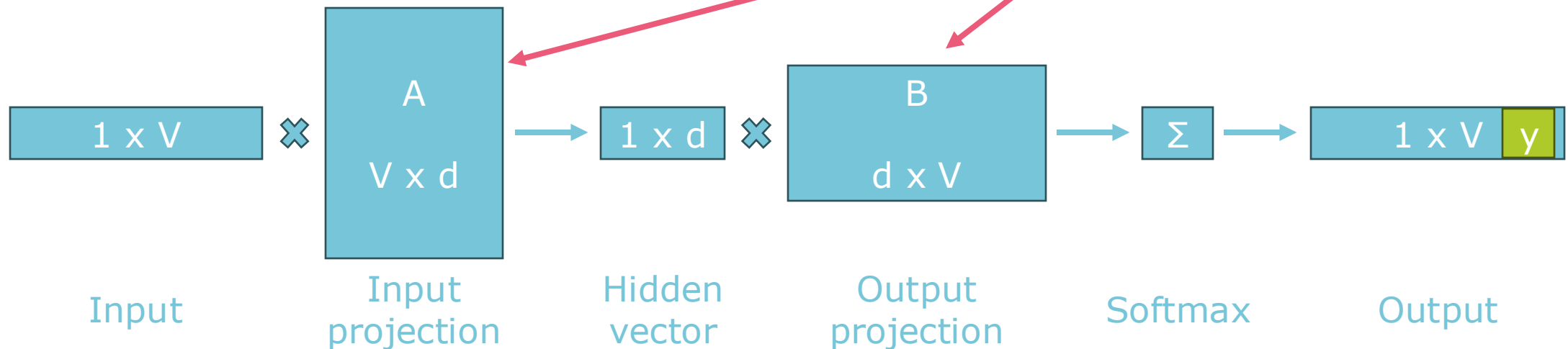


Word2vec Architecture

- 2-layer neural network
- Input: one-hot word vector, V : vocabulary size
- A: Projection to inner dimension d
- B: Projection to output
- Softmax Σ to predict target word y

Can use both as
word embeddings

(input projection
used more often,
empirical question)



Word2vec

Two variants:

- CBOW: Predict the center word from its context

the cat ____ on the

- Skip-gram: Predict the context words from the center word

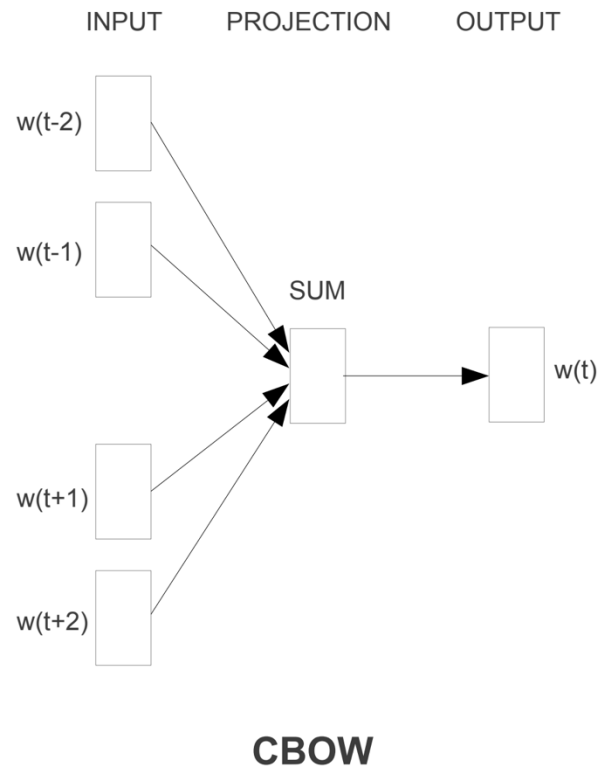
____ sat ____

Aside: Context Windows

Window Size	Text	Skip-grams
2	[The wide road shimmered] in the hot sun.	wide, the wide, road wide, shimmered
	The [wide road shimmered in the] hot sun.	shimmered, wide shimmered, road shimmered, in shimmered, the
	The wide road shimmered in [the hot sun].	sun, the sun, hot
3	[The wide road shimmered in] the hot sun.	wide, the wide, road wide, shimmered wide, in
	[The wide road shimmered in the hot] sun.	shimmered, the shimmered, wide shimmered, road shimmered, in shimmered, the shimmered, hot
	The wide road shimmered [in the hot sun].	sun, in sun, the sun, hot

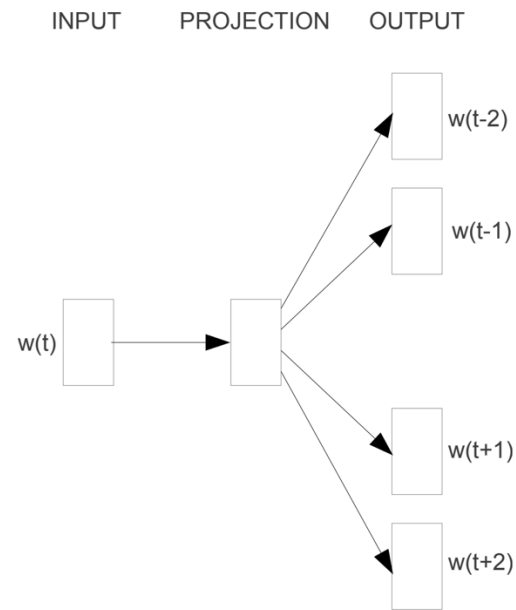
Continuous Bag-of-Words Model (CBOW)

- Predict the target center word from (the sum of) its context



Skip-gram Model

- Predict the context words from the center word
- 1 training example for each context word
 - $w(t) \rightarrow w(t-2)$
 - $w(t) \rightarrow w(t-1)$
 - ...



Skip-gram

Skip-gram Objective

- Objective: Increase log-likelihood of correct context word w_O given center word w_I :

$$p(w_O|w_I) = \frac{\exp \left(v'_{w_O}{}^\top v_{w_I} \right)}{\sum_{w=1}^W \exp \left(v'_w{}^\top v_{w_I} \right)}$$

- Infeasible computation in denominator: sum over all words in vocab for each training step

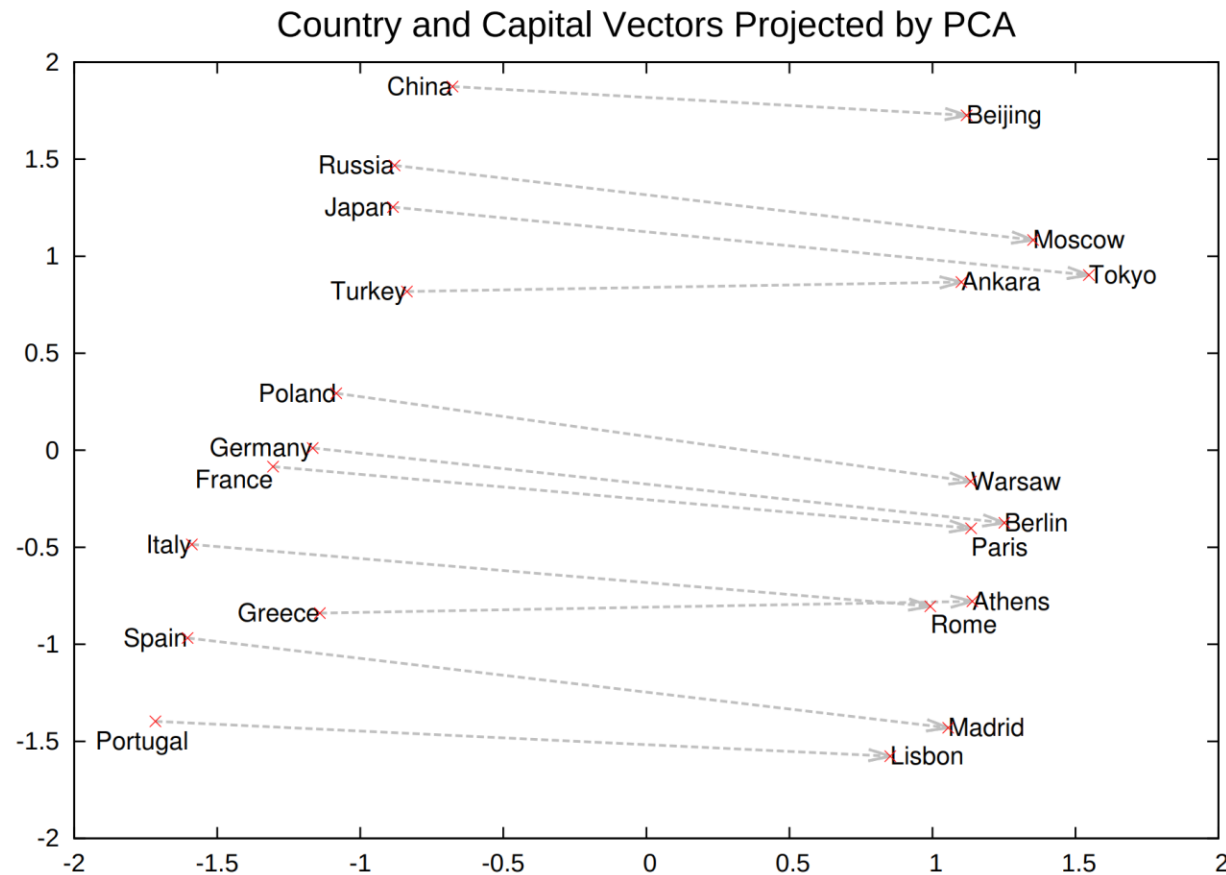
Skip-gram with Negative Sampling

- Idea: Negative Sampling
 - Sample some "distractor" (negative) words
 - Push probability of true (positive) word higher, negative words lower
 - One possible noise contrastive estimation (NCE) technique

$$\log \sigma(v'_{w_O}{}^\top v_{w_I}) + \sum_{i=1}^k \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-v'_{w_i}{}^\top v_{w_I}) \right]$$

- k negative examples for each positive
- $P_n(w)$: sample most frequent words less often

Structure in Embeddings

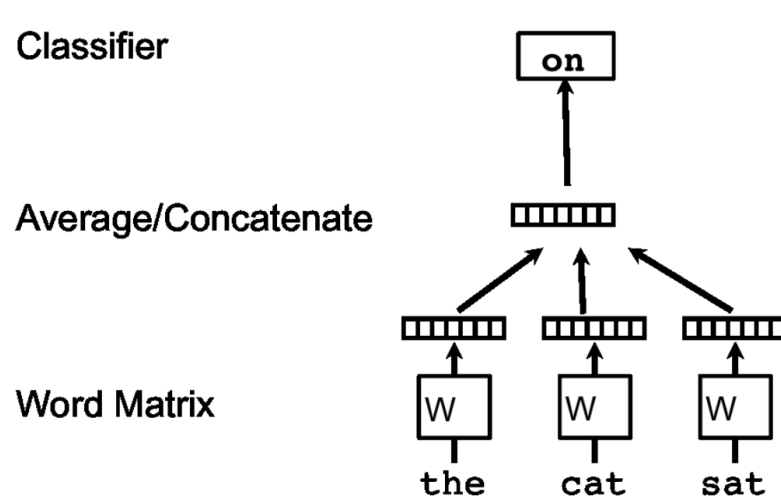


Extension: Phrase Embeddings

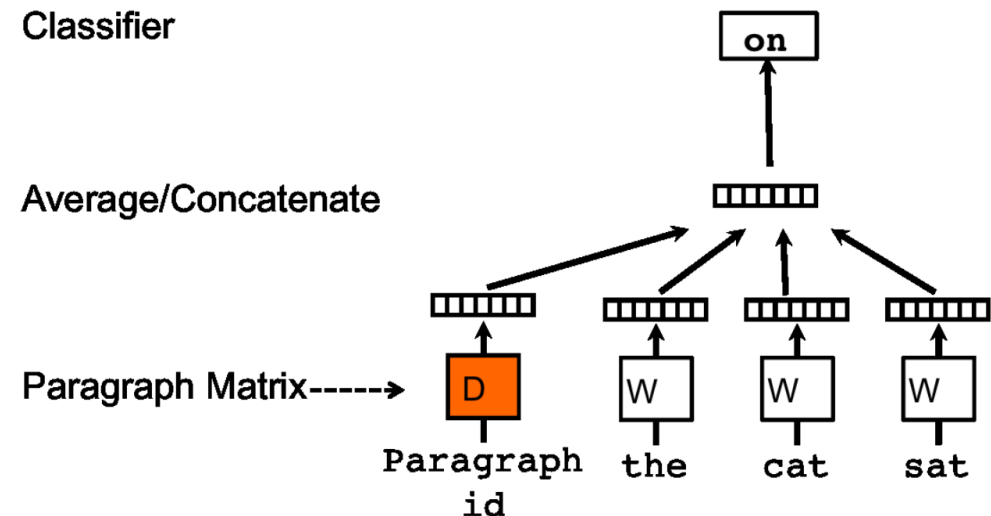
- Find words that appear frequently together, and infrequently in other contexts
 - "New York Times"
 - "Toronto Maple Leafs"
 - "this is"
- Replace them with a single token when they appear together
 - New York Times → New_York_Times

Extension: Doc2vec

- [Le & Mikolov, 2014](#)
- Add paragraph vector to center word prediction



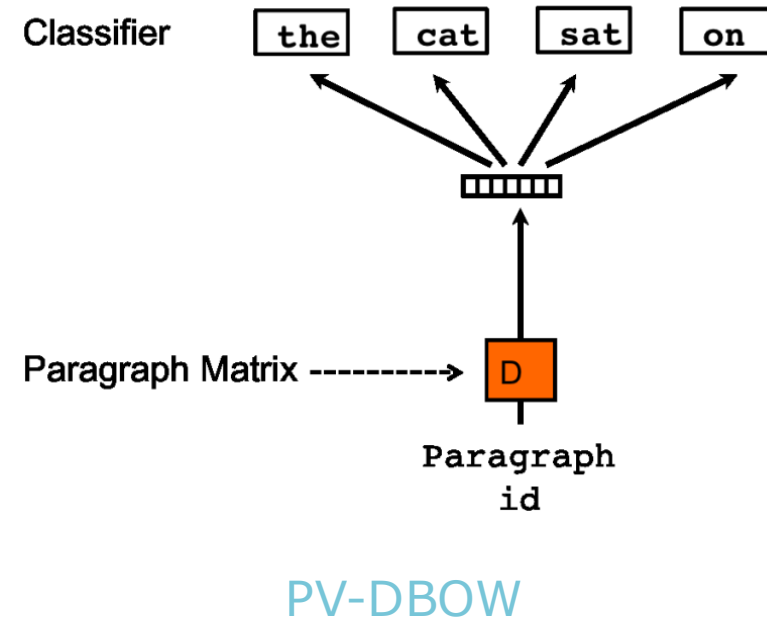
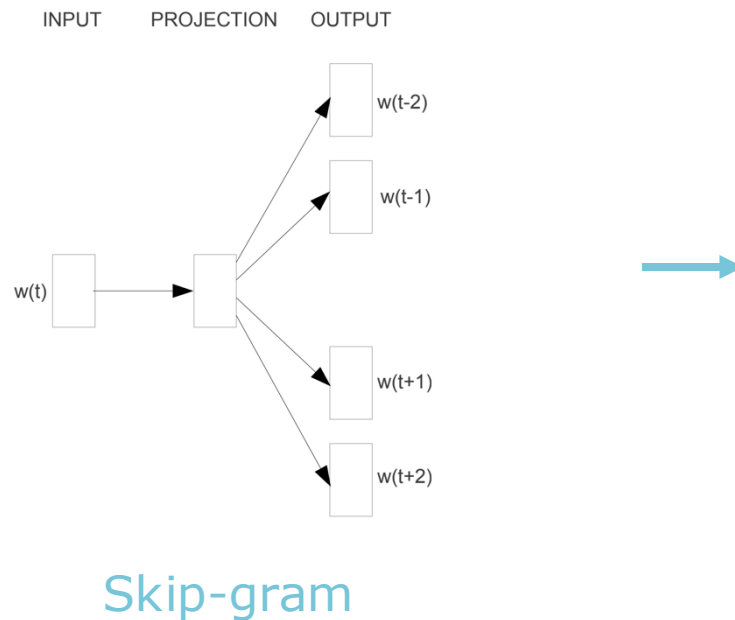
CBOW



PV-DM
(paragraph vectors with
distributed memory)

Extension: Doc2vec

- Also a version for skip-gram (with the unfortunate naming PV-DBOW, paragraph vectors with distributed bag of words)



GloVe

- [Pennington et al., 2014](#)
- Global Vectors (GloVe) for Word Representation
- Idea: Instead of negative sampling, train directly on the word co-occurrence matrix
- Word co-occurrence statistics is necessary to get rid of noise:

Table 1: Co-occurrence probabilities for target words *ice* and *steam* with selected context words from a 6 billion token corpus. Only in the ratio does noise from non-discriminative words like *water* and *fashion* cancel out, so that large values (much greater than 1) correlate well with properties specific to ice, and small values (much less than 1) correlate well with properties specific of steam.

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

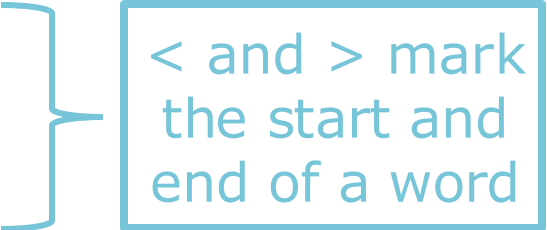
GloVe

- Least squares regression of word vectors w and \tilde{w}
 - Same distinction as in word2vec
 - They are the same if co-occurrence matrix X is symmetric

$$J = \sum_{i,j=1}^V f(X_{ij}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2, \quad (8)$$

- Down-weight rare co-occurrences (noisy) with function f
- Generally performs as well as word2vec, choice is personal preference (or empirical question)

fastText

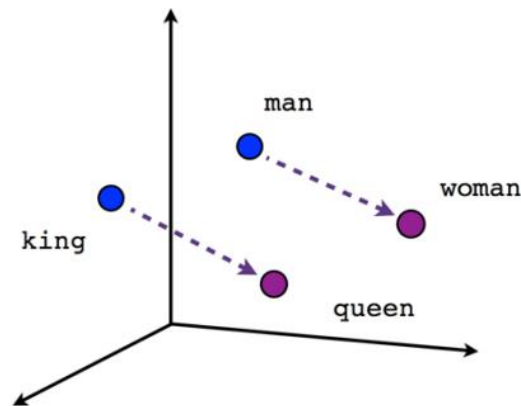
- [Bojanowski et al., 2016](#) & [Joulin et al., 2016](#)
 - Adds character n-gram embeddings to skip-gram idea
 - Bag of character n-grams ($3 \leq n \leq 6$) for each word...
 - E.g. $n=3$: where \rightarrow <wh, whe, her, ere, re>
 - ... and full word
 - <where>
- 
- < and > mark the start and end of a word
- Word representation is sum of bag of n-grams
 - Some n-grams more important (= meaningful) than others
 - Works especially well on morphologically rich languages and compounds ("Tischtennis")
 - Can handle unknown words (uses the sum of character n-grams)

Evaluation

- Intrinsic evaluation
 - Evaluates the word representations directly
 - Do they contain useful structure?
- Extrinsic evaluation
 - Evaluates the representations as input to a model (e.g. a classifier)
 - Do they improve performance on downstream tasks vs. one-hot/TF-IDF vectors?
 - As initialization of the embedding matrix in a neural network (this is no longer done, embeddings are trained from scratch)


Intrinsic Evaluation: Word Analogy

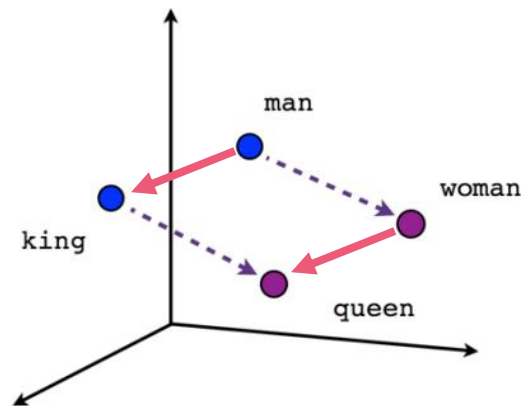
- Word embeddings contain structure for various relations
- Analogy task: “Man is to king, as woman is to what?”
- Compute $(\text{king} - \text{man} + \text{woman})$ and find closest* vector
 - *: Typically need to ignore other 3 vectors



Male-Female

Intrinsic Evaluation: Word Analogy

- Analogy task: "Man is to king, as woman is to what?"
- Alternative visualization: Apply man-to-king-vector  to woman



Male-Female

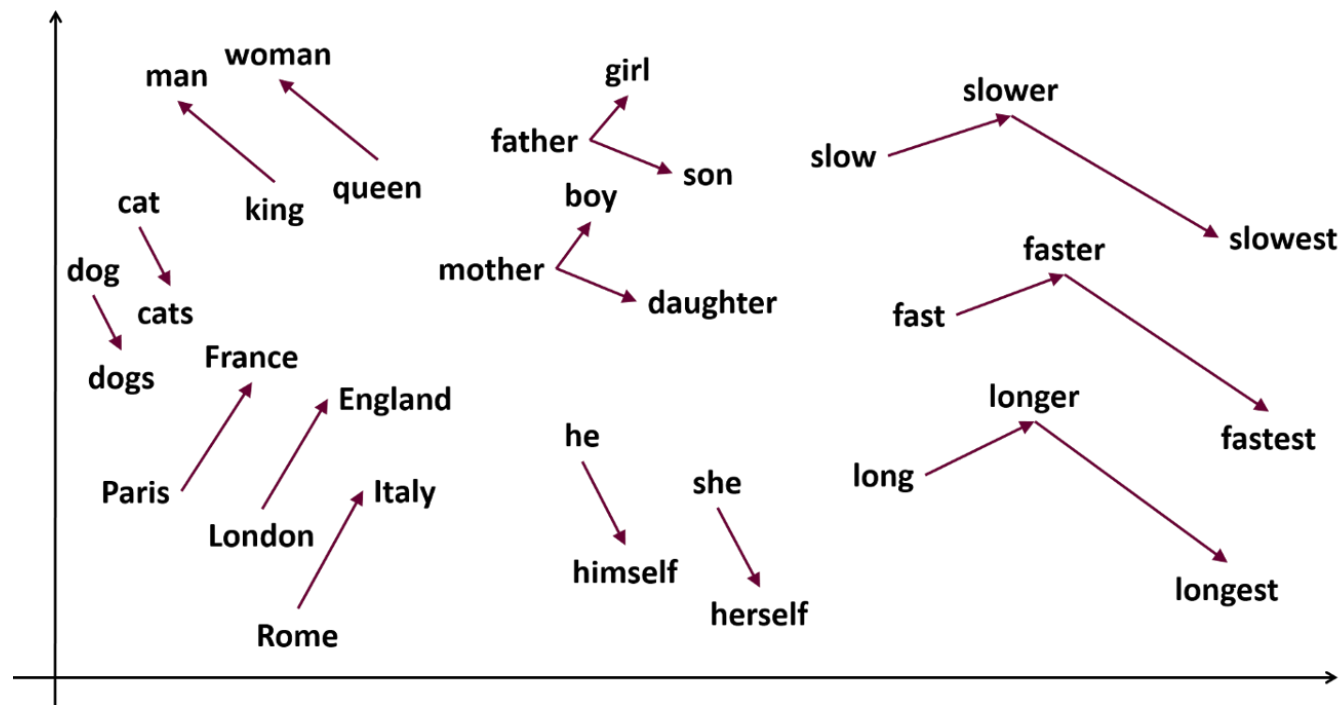
Word Analogy

- Vector arithmetic in word2vec
 - Athens (Greece, capital city) – Greece \approx vector for capital city
 - Athens – Greece + Norway \approx Oslo (Norway, capital city)

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

Word Analogy

- Can detect vector for a syntactic relation
 - Plural
 - Superlative
 - ...



Phrase Analogies

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

Interesting Analogies

- house:roof::castle:[dome, bell_tower, spire, crenellations, turrets]
- knee:leg::elbow:[forearm, arm, ulna_bone]
- New York Times:Sulzberger::Fox:[Murdoch, Chernin, Bancroft, Ailes]
 - The Sulzberger-Ochs family owns and runs the NYT.
 - The Murdoch family owns News Corp., which owns Fox News.
 - Peter Chernin was News Corp.'s COO for 13 yrs.
 - Roger Ailes is president of Fox News.
 - The Bancroft family sold the Wall St. Journal to News Corp.
- love:indifference::fear:[apathy, callousness, timidity, helplessness, inaction]
- Donald Trump:Republican::Barack Obama:[Democratic, GOP, Democrats, McCain]
- monkey:human::dinosaur:[fossil, fossilized, Ice_Age_mammals, fossilization]
 - Humans are what's left over from monkeys?
 - Monkeys evolved into humans, dinosaurs evolved into fossils?
- building:architect::software:[programmer, SecurityCenter, WinPcap]

Vector Arithmetic in Word2vec

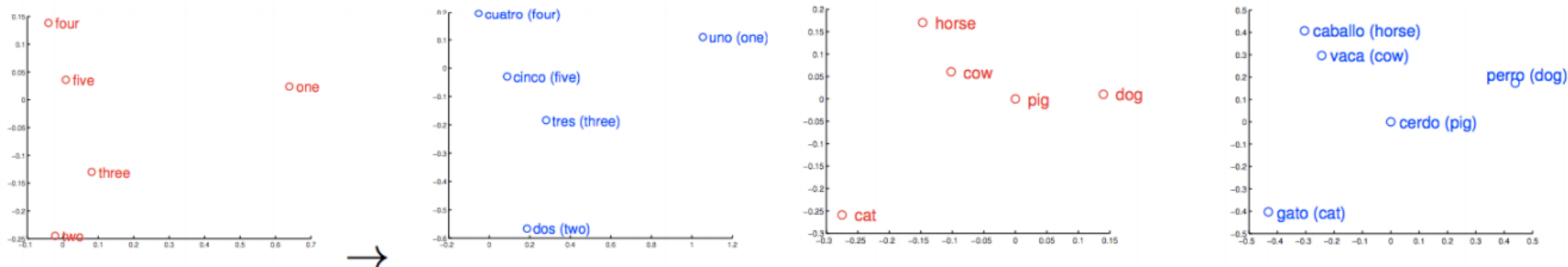
- Geopolitics: Iraq - Violence = Jordan
- Distinction: Human - Animal = Ethics
- President - Power = Prime Minister
- Library - Books = Hall
- Analogy: Stock Market \approx Thermometer

Word Similarity

- Many datasets, such as WordSim353/WS-353, Rare Words (RW), RG, SimLex-999, ...
- Ask humans to rate similarity of two words
- Compare cosine similarity of the two embeddings with human similarity rating
 - Higher correlation with human judgements is better
- Similarity vs. relatedness
 - Cup vs. coffee (related, but not similar)
 - Car vs. train (similar)
 - But: Similarity judgement is higher for cup & coffee than for car & train ([Faruqui et al., 2016](#))

Extrinsic Evaluation

- Document classification (topics, similarity, sentiment)
 - Average word vectors of a document
 - Train a classifier that takes averaged vector as input
- Translation: Find closest word in other language
 - Directly from a multilingual embedding (nearest neighbor)
 - From two monolingual embeddings, with a learned linear mapping from one space to the other



Exercise: Word Embeddings