

# WORD EMBEDDINGS

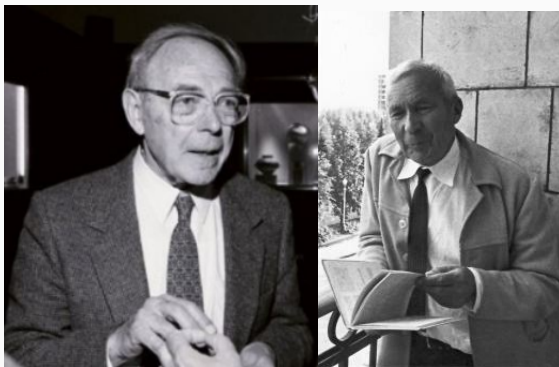
---

David Talbot

23rd April, 2017

Computer Science Club, St. Petersburg, Russia

# distributional hypothesis (harris 1957)



- Harris: Words which are *similar* occur in *similar contexts*.
- Kolmogorov: Defined *grammatical case* as set of contexts.

## metric space for nlp data?

- Some data comes with a natural associated metric
- E.g. Real numbers:  $d(x, y) = |y - x|$
- How about image data?

## metric space for nlp data?

- Some data comes with a natural associated metric
- E.g. Real numbers:  $d(x, y) = |y - x|$
- How about image data?
- How about speech?

## metric space for nlp data?

- Some data comes with a natural associated metric
- E.g. Real numbers:  $d(x, y) = |y - x|$
- How about image data?
- How about speech?
- How about natural language text?

## defining a *metric space* for language

Given vocab  $V$ , induce a 'distance'  $d : V \times V \rightarrow \mathbb{R}$

- Approach: Use distribution over auxiliary variable  $y$

$$d(w, w') = KL(\Pr(y|w) || \Pr(y|w')) = \sum_{y'} \Pr(y'|w) \log \frac{\Pr(y'|w)}{\Pr(y'|w')}$$

## maximum likelihood clustering (brown)

- Partition vocab  $V$  into  $G$  word classes  $C : V \rightarrow [0, G)$
- Model data as

$$\Pr(w_t | w_{t-1}) \approx \Pr(w_t | c_{w_t}) \Pr(c_{w_t} | c_{w_{t-1}})$$

- Maximize the loglikelihood of data under this model w.r.t.  $C$

## maximum likelihood clustering (brown)

Maximize the loglikelihood of data under this model w.r.t.  $C$

$$\ell(C) = \sum_{t=1}^T \log \Pr(w_t | w_{t-1}, w_{t-2}, \dots)$$



## maximum likelihood clustering (brown)

Maximize the loglikelihood of data under this model w.r.t.  $C$

$$\begin{aligned}\ell(C) &= \sum_{t=1}^T \log \Pr(w_t | w_{t-1}, w_{t-2}, \dots) \\ &\approx \sum_{(w, w') \in V^2} \mathbf{N}(w, w') \log \Pr(w | c_{w'}) \Pr(c_w | c_{w'})\end{aligned}$$

## maximum likelihood clustering (brown)

Maximize the loglikelihood of data under this model w.r.t.  $C$

$$\begin{aligned}\ell(C) &= \sum_{t=1}^T \log \Pr(w_t | w_{t-1}, w_{t-2}, \dots) \\ &\approx \sum_{(w, w') \in V^2} \mathbf{N}(w, w') \log \Pr(w | c_{w'}) \Pr(c_w | c_{w'}) \\ &= \sum_{c_w, c_{w'}} \mathbf{N}(c_w, c_{w'}) \log \frac{\mathbf{N}(c_w, c_{w'})}{\mathbf{N}(c_w) \mathbf{N}(c_{w'})} + \sum_w \mathbf{N}(w) \log \frac{\mathbf{N}(w)}{\mathbf{N}(c_w)}\end{aligned}$$

# maximum likelihood clustering (brown)

Maximize the loglikelihood of data under this model w.r.t.  $C$

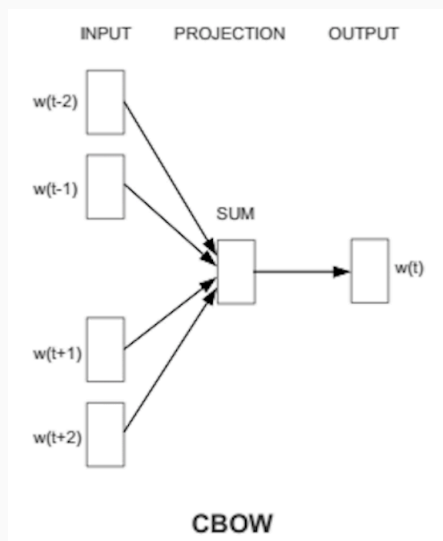
$$\begin{aligned}\ell(C) &= \sum_{t=1}^T \log \Pr(w_t | w_{t-1}, w_{t-2}, \dots) \\ &\approx \sum_{(w, w') \in V^2} \mathbf{N}(w, w') \log \Pr(w | c_{w'}) \Pr(c_w | c_{w'}) \\ &= \sum_{c_w, c_{w'}} \mathbf{N}(c_w, c_{w'}) \log \frac{\mathbf{N}(c_w, c_{w'})}{\mathbf{N}(c_w) \mathbf{N}(c_{w'})} + \sum_w \mathbf{N}(w) \log \frac{\mathbf{N}(w)}{\mathbf{N}(c_w)} \\ &= \mathbf{I}(c_w, c'_w) + \mathbf{H}(c_w) - \mathbf{H}(w)\end{aligned}$$

where  $\mathbf{N}(w)$  and  $\mathbf{N}(w, w')$  are counts of  $w$  and  $(w, w')$  respectively.

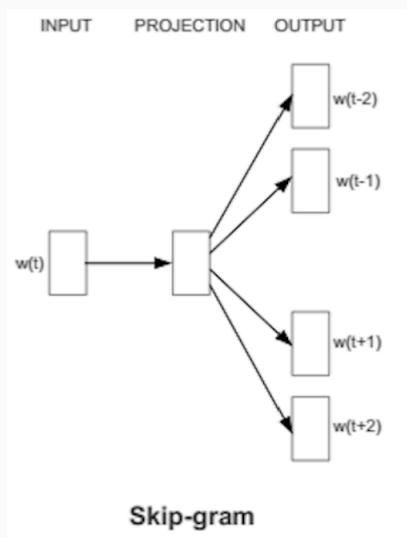
## distributed continuous representations

- Explicit representation e.g.  $\Pr(y|x)$  is sparse
- Embedding into lower-dimensional vector, e.g. *word2vec*

## word2vec models (mikolov 2013)



## word2vec models (mikolov 2013)



## skip-gram model details (goldberg & levy 2014)

Maximize the loglikelihood of data under the skip-gram model  
w.r.t. embedding  $\theta$

$$\arg \max_{\theta} = \prod_{(w,c) \in D} \Pr(c|w; \theta)$$

which is parameterized as

$$\Pr(c|w; \theta) = \frac{\exp(v_c \cdot v_w)}{\sum_{c' \in \mathcal{C}} \exp(v_{c'} \cdot v_w)}$$

where  $v_c, v_w \in \mathbb{R}^d$ .

## skip-gram model details

Which is equivalent to

$$\arg \max_{\theta} \sum_{(w,c) \in D} \log \Pr(c|w; \theta) = \sum_{(w,c) \in D} (e^{v_c \cdot v_w} - \log \sum_{c'} e^{v_{c'} \cdot v_w})$$



## skip-gram model details

Which is equivalent to

$$\arg \max_{\theta} \sum_{(w,c) \in D} \log \Pr(c|w; \theta) = \sum_{(w,c) \in D} (e^{(v_c \cdot v_w)} - \log \sum_{c'} e^{(v_{c'} \cdot v_w)})$$

- Use negative sampling to approximate sum over  $c'$
- Forces model to discriminate observed data from noise

- LSA (Latent Semantic Analysis): Apply SVD to count matrix  $M$

$$M \approx \hat{M}_d = W_d \Sigma_d C_d.$$

## other sparse embeddings

- LSA (Latent Semantic Analysis): Apply SVD to count matrix  $M$

$$M \approx \hat{M}_d = W_d \Sigma_d C_d.$$

- GloVe: factorize shifted log-count matrix

$$v_w \cdot v_c + b_w + b_c = \log(\#(w, c)) \quad \forall (w, c) \in D.$$

## other sparse embeddings

- LSA (Latent Semantic Analysis): Apply SVD to count matrix  $M$

$$M \approx \hat{M}_d = W_d \Sigma_d C_d.$$

- GloVe: factorize shifted log-count matrix

$$v_w \cdot v_c + b_w + b_c = \log(\#(w, c)) \quad \forall (w, c) \in D.$$

What will vector differences look like for GloVe?

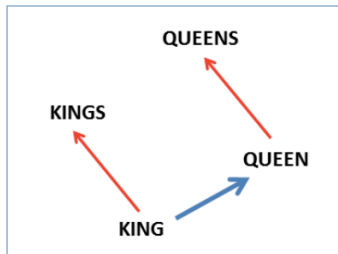
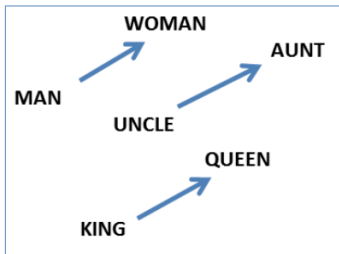
## glove: conditional ratios (pennington et al. 2014)

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

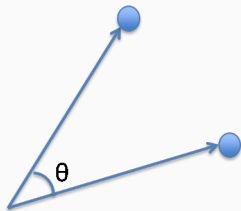
GloVe vector differences approximate logarithm of their ratios

$$v_x \cdot v_c \approx \log(\#(x, c)) \implies |v_x - v_y| \cdot v_c \approx \log \frac{\Pr(x|c)}{\Pr(y|c)}.$$

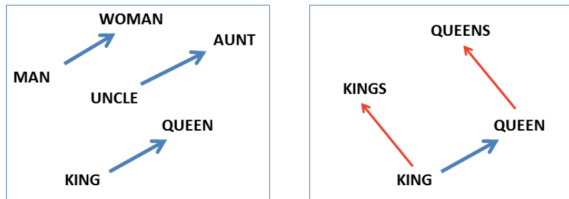
## vector offsets between word embeddings (mikolov 2013)



$$\text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$



# analogical reasoning in vector space (mikolov 2013)



Syntactic relations (e.g. morphology)

apples – apple  $\approx$  cars – car

Semantic relations

queen – woman  $\approx$  king – man



## alternative search objectives (goldberg & levy 2014)

Given  $(x, x', y)$  find  $y' \in V$  that maximizes:

$$\text{Cos3Add} = \arg \max_{y' \in V} (\cos(y', y - x + x'))$$

## alternative search objectives (goldberg & levy 2014)

Given  $(x, x', y)$  find  $y' \in V$  that maximizes:

$$\text{Cos3Add} = \arg \max_{y' \in V} (\cos(y', y - x + x'))$$

Alternative that preserves *direction* of transformation

$$\text{PairDirections} = \arg \max_{y' \in V} (\cos(y' - y, x' - x))$$

## alternative search objectives (goldberg & levy 2014)

Given  $(x, x', y)$  find  $y' \in V$  that maximizes:

$$\text{Cos3Add} = \arg \max_{y' \in V} (\cos(y', y - x + x'))$$

Alternative that preserves *direction* of transformation

$$\text{PairDirections} = \arg \max_{y' \in V} (\cos(y' - y, x' - x))$$

If vectors are normalized, then the first can be written:

$$\arg \max_{y' \in V} (\cos(y', y) + \cos(y', x') - \cos(y', x))$$

## how well do embeddings perform?

<b>Representation</b>	<b>MSR</b>	<b>GOOGLE</b>	<b>SEMEVAL</b>
Embedding	53.98%	62.70%	38.49%
Explicit	29.04%	45.05%	38.54%

Table 1: Performance of **3COSADD** on different tasks with the explicit and neural embedding representations.

<b>Representation</b>	<b>MSR</b>	<b>GOOGLE</b>	<b>SEMEVAL</b>
Embedding	9.26%	14.51%	44.77%
Explicit	0.66%	0.75%	45.19%

Table 2: Performance of **PAIRDIRECTION** on different tasks with the explicit and neural embedding representations.

## problems with cos3add (goldberg & levy 2014)

Soft-OR behaviour: one sufficiently large term can dominate

$$\arg \max_{y' \in V} (\cos(y', y) + \cos(y', x') - \cos(y', x))$$

For example

$$\arg \max_{y' \in V} (\cos(y', \text{Baghdad}) + \cos(y', \text{England}) - \cos(y', \text{London}))$$

Returns *Mosul* rather than *Iraq*

Proposed alternative (equivalent to taking logs):

$$3\text{CosMul} = \arg \max_{y' \in V} \frac{\cos(y', y) \cos(y', x')}{\cos(y', x) + \epsilon}$$

## how well do embeddings perform?

<b>Objective</b>	<b>Representation</b>	<b>MSR</b>	<b>GOOGLE</b>
3COSADD	Embedding	53.98%	62.70%
	Explicit	29.04%	45.05%
3COSMUL	Embedding	<b>59.09%</b>	66.72%
	Explicit	56.83%	<b>68.24%</b>

Table 3: Comparison of **3COSADD** and **3COSMUL**.

# how well do embeddings perform?

	Relation	Embedding	Explicit
GOOGLE	capital-common-countries	90.51%	<b>99.41%</b>
	capital-world	77.61%	<b>92.73%</b>
	city-in-state	56.95%	<b>64.69%</b>
	currency	<b>14.55%</b>	10.53%
	family (gender inflections)	<b>76.48%</b>	60.08%
	gram1-adjective-to-adverb	<b>24.29%</b>	14.01%
	gram2-opposite	<b>37.07%</b>	28.94%
	gram3-comparative	<b>86.11%</b>	77.85%
	gram4-superlative	56.72%	<b>63.45%</b>
	gram5-present-participle	63.35%	<b>65.06%</b>
	gram6-nationality-adjective	89.37%	<b>90.56%</b>
	gram7-past-tense	<b>65.83%</b>	48.85%
	gram8-plural (nouns)	72.15%	<b>76.05%</b>
	gram9-plural-verbs	<b>71.15%</b>	55.75%
MSR	adjectives	45.88%	<b>56.46%</b>
	nouns	56.96%	<b>63.07%</b>
	verbs	<b>69.90%</b>	52.97%

Table 5: Breakdown of relational similarities in each representation by relation type, using 3CosMUL.

## inferring morphological relations (soricut & och, 2015)

Given sets of word pairs that differ in a common edit

$(\text{suf} = \emptyset, \text{suf} = -s) = \{(\text{dog}, \text{dogs}), (\text{cat}, \text{cats}), \dots\}$

$(\text{suf} = -\text{ing}, \text{suf} = -\text{ed}) = \{(\text{playing}, \text{played}), (\text{walking}, \text{walked}), \dots\}$

$(\text{pref} = r-, \text{pref} = \text{str}-) = \{(\text{ring}, \text{string}), (\text{rayed}, \text{strayed}), \dots\}$

Use vector space of embeddings to find *valid* transformations



Evaluate transformation  $r$  e.g. (suf = -ing, suf = -ed)

$$r : w \in V \rightarrow w' \in V$$

Evaluate transformation  $r$  e.g. (suf = -ing, suf = -ed)

$$r : w \in V \rightarrow w' \in V$$

Define  $S_r$  as

$$S_r = (w, w') \in V^2 \quad \text{s.t.} \quad r(w) = w'$$

Evaluate transformation  $r$  e.g. (suf = -ing, suf = -ed)

$$r : w \in V \rightarrow w' \in V$$

Define  $S_r$  as

$$S_r = (w, w') \in V^2 \quad \text{s.t.} \quad r(w) = w'$$

Evaluate each pair  $(w_1, w_2) \in S_r$  against all others  $(w, w') \in S_r$

## inferring morphological relations (soricut & och, 2015)

Evaluate transformation  $r$  e.g. (suf = -ing, suf = -ed)

$$r : w \in V \rightarrow w' \in V$$

Define  $S_r$  as

$$S_r = (w, w') \in V^2 \quad \text{s.t.} \quad r(w) = w'$$

Evaluate each pair  $(w_1, w_2) \in S_r$  against all others  $(w, w') \in S_r$

$$Eval\{(w_1, w_2)\} = \frac{1}{|S_r|} \sum_{(w, w') \in S_r} rank_{cos}(w_2, w_1 - w + w').$$

## inferring morphological relations (soricut & och, 2015)

Evaluate transformation  $r$  e.g. ( $\text{suf} = -\text{ing}$ ,  $\text{suf} = -\text{ed}$ )

$$r : w \in V \rightarrow w' \in V$$

Define  $S_r$  as

$$S_r = (w, w') \in V^2 \quad \text{s.t.} \quad r(w) = w'$$

Evaluate each pair  $(w_1, w_2) \in S_r$  against all others  $(w, w') \in S_r$

$$\text{Eval}\{(w_1, w_2)\} = \frac{1}{|S_r|} \sum_{(w, w') \in S_r} \text{rank}_{\cos}(w_2, w_1 - w + w').$$

How about ambiguous rules such as ( $\text{suf} = \emptyset$ ,  $\text{suf} = -\text{s}$ )?