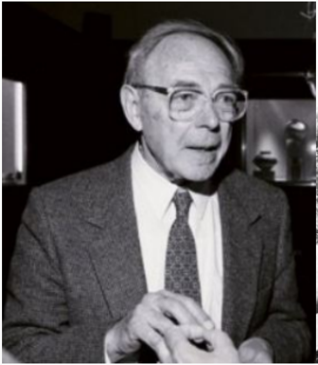
WORD EMBEDDINGS

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' $d : V \times V \to \mathbb{R}$

- 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 $C : V \to [0, G)$
- 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 loglikehood of data under this model w.r.t. C

maximum likelihood clustering (brown)

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

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

maximum likelihood clustering (brown)

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

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

maximum likelihood clustering (brown)

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

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

maximum likelihood clustering (brown)

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

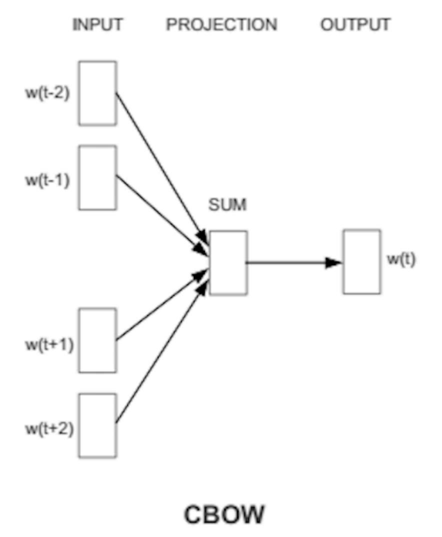
$$
\begin{aligned}
\ell(C) &= \sum_{t=1}^{T} \log \mathsf{Pr}(w_t | w_{t-1}, w_{t-2}, \dots) \\
&\approx \sum_{(w,w') \in V^2} \mathsf{N}(w, w') \log \mathsf{Pr}(w | c_{w'}) \mathsf{Pr}(c_w | c_{w'}) \\
&= \sum_{c_w, c_{w'}} \mathsf{N}(c_w, c_{w'}) \log \frac{\mathsf{N}(c_w, c_{w'})}{\mathsf{N}(c_w)\mathsf{N}(c_{w'})}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

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

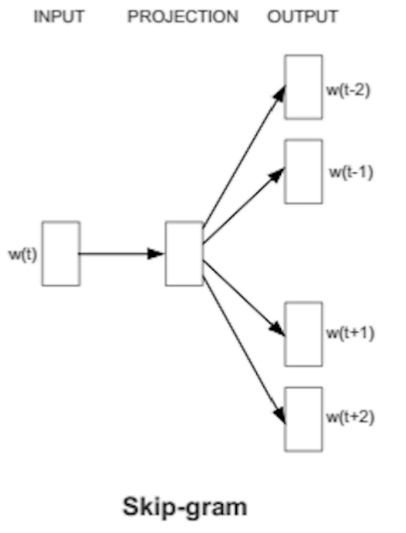
distributed continuous representations

- 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 loglikehood of data under the skip-gram mode w.r.t. embedding θ

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

which is parameterized as

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

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

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

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

other sparse embeddings

- 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?

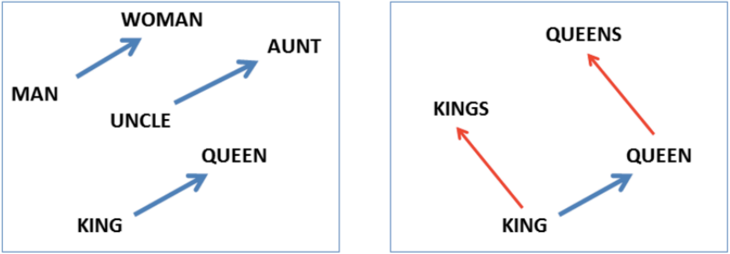
glove: conditional ratios (pennington et al. 2014)

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

GloVe vector differences approximate logarithm of their ratios

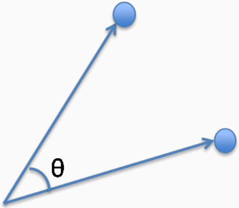
$$v_x \cdot v_c \approx \log(\#(x, c)) \implies |v_x - v_y| \cdot v_c \approx \log \frac{\Pr(x|c)}{\Pr(y|c)}.$$

vector offsets between word embeddings (mikolov 2013)

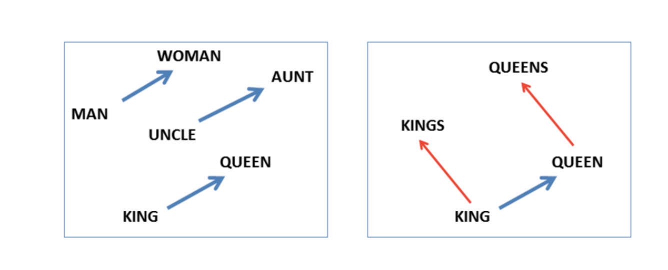


cosine similarity

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



analogical reasoning in vector space (mikolov 2013)



Syntactic relations (e.g. morphology)

$$\text{apples} - \text{apple} \approx \text{cars} - \text{car}$$

Semantic relations

$$\text{queen} - \text{woman} \approx \text{king} - \text{man}$$

alternative search objectives (goldberg & levy 2014)

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

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

alternative search objectives (goldberg & levy 2014)

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

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

Alternative that preserves *direction* of transformation

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

alternative search objectives (goldberg & levy 2014)

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

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

Alternative that preserves *direction* of transformation

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

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?

Representation	MSR	GOOGLE	SEMEVAL
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.

Representation	MSR	GOOGLE	SEMEVAL
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

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

Returns *Mosul* rather than *Iraq*

Proposed alternative (equivalent to taking logs):

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

how well do embeddings perform?

Objective	Representation	MSR	GOOGLE
3COSADD	Embedding	53.98%	62.70%
	Explicit	29.04%	45.05%
3COSMUL	Embedding	**59.09%**	66.72%
	Explicit	56.83%	**68.24%**

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

how well do embeddings perform?

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

$$
\begin{aligned}
(\text{suf} = \emptyset, \text{suf} = -\text{s}) &= \{(\text{dog}, \text{dogs}), (\text{cat}, \text{cats}), \ldots\} \\
(\text{suf} = -\text{ing}, \text{suf} = -\text{ed}) &= \{(\text{playing}, \text{played}), (\text{walking}, \text{walked}), \ldots\} \\
(\text{pref} = \text{r}-, \text{pref} = \text{str}-) &= \{(\text{ring}, \text{string}), (\text{rayed}, \text{strayed}), \ldots\}
\end{aligned}
$$

Use vector space of embeddings to find *valid* transformations

inferring morphological relations (soricut & och, 2015)

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

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

inferring morphological relations (soricut & och, 2015)

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

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

Define S_r as

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

inferring morphological relations (soricut & och, 2015)

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

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

Define S_r as

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

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. $(\mathrm{suf} = -\mathrm{ing}, \mathrm{suf} = -\mathrm{ed})$

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

Define S_r as

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

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. $(\mathrm{suf} = -\mathrm{ing}, \mathrm{suf} = -\mathrm{ed})$

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

Define S_r as

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

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

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

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