# Distributional Models for Lexical Semantics 

## Lecture 1: Don't be afraid of vectors

Denis Paperno TyLex - Voronovo 2017

## Non-technical definition

- Distributional semantic models
- are ways of creating lexical semantic representations
- through learning by association
- on a large scale


# Distributional Semantic Models are vector spaces built from distributional information 

Vector: $\left(a_{1}, a_{2}, \ldots, a_{n}\right)$, where each $a_{i}$ is an element of $F$.

## Origins

- It may be presumed that any two morphemes $A$ and $B$ having different meanings, also differ somewhere in distribution: there are some environments in which one occurs and the other does not (Harris 1951)
- "You shall know a word by the company it keeps." (Firth 1957)


## Example

he curtains open and the moon shining in on the barely ars and the cold , close moon " . And neither of the w rough the night with the moon shining so brightly, it made in the light of the moon . It all boils down , wr surely under a crescent moon , thrilled by ice-white sun, the seasons of the moon ? Home, alone, Jay pla $m$ is dazzling snow, the moon has risen full and cold un and the temple of the moon , driving out of the hug in the dark and now the moon rises , full and amber a bird on the shape of the moon over the trees in front But I could n't see the moon or the stars, only the rning , with a sliver of moon hanging among the stars they love the sun , the moon and the stars . None of the light of an enormous moon . The plash of flowing w man 's first step on the moon ; various exhibits , aer the inevitable piece of moon rock . Housing The Airsh oud obscured part of the moon . The Allied guns behind

## Co-occurrence vector space

- Simplest distributional model: just count!
- Vectors represent words
- Dimensions represent contexts

|  | bark | walk | talk | tail | bag |
| :--- | :--- | :--- | :--- | :--- | :--- |
| dog | 246 | 72 | 78 | 71 | 1 |
| cat | 5 | 15 | 25 | 32 | 0 |
| man | 0 | 57 | 133 | 0 | 1 |
| moman | 2 | 203 | 407 | 5 | 18 |

## Vectors as word representations



## Vector similarity measures

$\cos (a, b)=\frac{\sum_{i} a_{i} * b_{i}}{\sqrt{\left(\sum_{i} a_{i}^{2}\right) *\left(\sum_{i} b_{i}^{2}\right)}}$

$$
\operatorname{euc}(a, b)=\sqrt{\sum_{i}\left(a_{i}-b_{i}\right)^{2}}
$$

## Vectors as word representations



## Cosine values

- 1 for identical vectors
- 0 for orthogonal vectors
- Negative values rare for linguistic vectors


## Example



## Example

## $\cos (a, b)=\quad \sum_{i} a_{i} * b_{i}$ <br> 

- cos(cat,man)=
$5 * 0+15 * 57+25 * 133+32 * 0+0 * 1$
$\sqrt{\left(5^{2}+15^{2}+25^{2}+32^{2}+0^{2}\right)\left(0^{2}+57^{2}+133^{2}+0^{2}+1^{2}\right)}$

| cat | 5 | 15 | 25 | 32 | 0 |
| :--- | :--- | :--- | :--- | :--- | :--- |
| man | 0 | 57 | 133 | 0 | 1 |

## Example

## $\cos (a, b)=\quad \sum_{i} a_{i} * b_{i}$ <br> $\cos (a, b)=$ <br> $$
=\sqrt{\sqrt{\left(\sum_{i}^{2} a_{i}^{2}\right) *}\left(\sum_{i} b_{i}^{2}\right)}
$$

- $\cos (c a t, m a n)=$
$5 * 0+15 * 57+25 * 133+32 * 0+0 * 1$
$\sqrt{\left(5^{2}+15^{2}+25^{2}+32^{2}+0^{2}\right)\left(0^{2}+57^{2}+133^{2}+0^{2}+1^{2}\right)}$
$=4180 / \sqrt{ } 1899 * 20939 \approx .66$

| cat | 5 | 15 | 25 | 32 | 0 |
| :--- | :--- | :--- | :--- | :--- | :--- |
| man | 0 | 57 | 133 | 0 | 1 |

## Application: similarity and relatedness

Examples from WordSim-353:

| Word pair | relatedness | cosine |
| :--- | :--- | :--- |
| money~cash | 9.15 | .98 |
| tiger~zoo | 5.87 | .42 |
| stock~phone | 1.62 | .04 |

Similarity and relatedness datasets exist of other languages, including Russian (Panchenko et al. 2016)

## Weighting

## the , (comma) be owner walk <br> dog $51764628 \quad 3195 \quad 245 \quad 237$

## Parameters of DSM

- weighting

$$
P M I(w, c)=\log \frac{\hat{P}(w, c)}{\hat{P}(w) \cdot \hat{P}_{\alpha}(c)}
$$

$P P M I(w, c)=\max \left(\log \frac{P(w, c)}{P(w) P(c)}, 0\right)$
$S P P M I_{k}(w, c)=\max (P M I(w, c)-\log k, 0)$

## PMI Weighting

## the , (comma) be owner walk dog $51764628 \quad 3195 \quad 245237$

the , (comma) be $\begin{array}{llllll}\operatorname{dog} & 1.6 & 1.52 & 1.56 & 3.05 & 2.73\end{array}$

## Application: Free Association

- What associations do you have when you hear the word PEN?


## Free association

Human associations:

- PEN
- PENCIL
- INK
- PAPER
- WRITE


## Free association

Human associations:

- PEN
- PENCIL
- INK
- PAPER
- WRITE

Distributional model:

- PEN
- PENCIL
- FOUNTAIN
- INK
- PAPER
- WRITE

Example from Griffiths et al. 2007, p. 223

## Application: semantic proportions

- man:king=woman:x
$\mathrm{x} \approx k i n g+w o m a n-m a n$
- Additive method:
( $\cos (\mathrm{x}$, king $)+\cos (\mathrm{x}$, woman $)-\cos (\mathrm{x}$, man $))$
- Multiplicative method:
$\cos (\mathrm{x}, \text { king })^{\star} \cos (\mathrm{x}$, woman)/cos(x,man)
(Levy and Goldberg 2014)


## Application: semantic proportions

- Works well for some relations (capital~country, gender) Russia:Moscow=Latvia:x
- Less well for others (currency~country, adverb~adjective)
Russia:ruble=Latvia: $x$

\author{

1. копейка 0.51 <br> 2. руб 0.50 <br> 3. лат 0.48 <br> 4. злотый 0.47 <br> 5. риксдалер 0.46
}

## Application: pragmatic alternatives

## Sentences of the form

- This is not an $X$, it is a $Y$.
- There is no X here but there is $\mathrm{a} Y$.


## Examples:

- This is not an alligator, this is a crocodile.
- There is no garlic here but there is a vampire.

Cosine (garlic,vampire) has high correlation with the plausibility of the sentence:

- . 86 for THIS IS,
- . 89 for THERE IS (Kruszewski et al. 2016)


## Toward a new application

- Assume similarity and relatedness solved; a general semantic task that further tests semantic models?
- Idea: test for semantic differences
- Semeval 2018 Task 10
https://competitions.codalab.org/competitions/17326

| Word 1 | Word 2 | feature | difference |
| :--- | :--- | :--- | :--- |
| dolphin | seal | (has) fins | YES |
| dolphin | seal | (eats) fish | NO |

## Have fun

- Visualize related English words http://www.serelex.org/ by Aleksandr Panchenko
- Vectors for Russian: http://rusvectores.org by Andrei Kutuzov and Elizaveta Kuzmenko

