Distributed Semantic Models for Word Vectors

COALS

Ramaseshan Ramachandran

1. COALS
2. First and Second Order associations
3. Dense Vectors
4. References

HAL SUMMARY

- A count-based word embedding model
- Captures word meanings through the unsupervised analysis of text
- Produces word vectors that are semantic (similar words) and associative in nature
- Acquires word meanings as a function of keeping track of how words are used in context
- Carries the history of the contextual experience by using a moving window and weighting of co-occurring words based on the distance
- Exploits the regularities of language such that conceptual generalisations can be captured in a data matrix

IMPACT OF FREQUENCY MEASURE ON SIMILARITY

- Even if t_1 and t_2 are unrelated, if $p(t_1) \approx p(t_2)$, then their vectors will contain elements with similar magnitudes.

  $\implies$ any similarity measure

  For example, words a, an, the co-occur with many words in the vocabulary

- Conversely if they are related but $p(t_1) \ll p(t_2)$ then their vectors will contain elements with widely differing magnitudes, simply due to their differing co-occurrence probability.

In general, relative frequency does not imply semantic similarity. Hence we require normalized measures ro build word vectors.

Correlated Occurrence Analogue to Lexical Semantic[1](COALS)

COALS METHODOLOGY

- Gather co-occurrence counts, typically ignoring closed-class neighbors and using a ramped window of size 4.
- Discard all but the m (14,000, in this case) columns reflecting the most common open-class words.
- Convert counts to word pair correlations - Instead of using the raw frequency score, correlation score is used to analyze the relationship between pair of words
- The correlation coefficient values with this normalization will be in the range of [-1,1]
- Set negative values to 0.
- Take square root of positive values.
- The semantic similarity between two words is given by the correlation of their vectors.
- The matrix constructed using this correlation would be semantic space
- COALS method employs a normalization strategy that largely factors out lexical frequency.
- Columns representing low-frequency words are removed

NORMALIZATION PROCEDURES

Several vector normalization procedures.

Row:	$w'_{a,b} = \frac{w_{a,b}}{\sum_j w_{a,j}}$
Column:	$w'_{a,b} = \frac{w_{a,b}}{\sum_i w_{i,b}}$
Length:	$w'_{a,b} = \frac{w_{a,b}}{(\sum_j {w_{a,j}}^2)^{1/2}}$
Correlation:	$w'_{a,b} = \frac{T w_{a,b} - \sum_j w_{a,j} \sum_i w_{i,b}}{\left(\sum_j w_{a,j} \cdot (T - \sum_j w_{a,j}) \cdot \sum_i w_{i,b} \cdot (T - \sum_i w_{i,b})\right)^{1/2}}$ $T = \sum_i \sum_j w_{i,j}$
Entropy:	$w'_{a,b} = \log(w_{a,b} + 1)/H_a$ $H_a = -\sum_j \frac{w_{a,b}}{\sum_j w_{a,j}} \log\left(\frac{w_{a,b}}{\sum_j w_{a,j}}\right)$

$$\mathrm{PMI}(w_i, w_c) = \log_2\left(\frac{p(w_i, w_c)}{p(w_i)p(w_c)}\right)$$

The range of PMI is $[-\infty, \infty]$. Positive PMI refers to word that often co-occur, while negative indicates that they are almost independent of each other or they co-occur less often. To focus only on similarity, we can consider only positive PMI as follows:

$$\mathrm{PPMI}(w_i, w_c) = \max\left(\log\left(\frac{p(w_i, w_c)}{p(w_i)p(w_c)}\right), 0\right)$$

BIAL IN NORMALIZATION

All normalization procedures have bias. PMI score is high for rare words. One way to reduce this bias toward low frequency words is to scale down the context count for PMI as follows:

$$\mathrm{PPMI}(w_i, w_c)_b = \max\left(\log\left(\frac{p(w_i, w_c)}{p(w_i)p_b(w_c)}\right), 0\right)$$

$$p_b(c) = \frac{\mathrm{count}(c)^b}{\sum_c \mathrm{count}(c)^b}$$

where $b \approx 0.75$

$b \approx 0.75$ increases the $p_b(c)$[2]. Hence $p_b(c) > p(c)$

1. Gather co-occurrence counts, typically ignoring closed-class neighbors and using a ramped, size 4 window:

   1 2 3 4 [0] 4 3 2 1

2. Discard all but the m (14,000, in this case) columns reflecting the most common open-class words.
3. Convert counts to word pair correlations, set negative values to 0, and take square roots of positive ones.
4. The semantic similarity between two words is given by the correlation of their vectors.

BUILDING CO-OCCURRENCE MATRIX - COALS STEP 1

How much wood would a woodchuck chuck, if a woodchuck could chuck wood? As much wood as a woodchuck would,if a woodchuck could chuck wood.

Step 1 of the COALS method: The initial co-occurrence table with a ramped, 4-word window.

	a	as	chuck	could	how	if	much	wood	woodch.	would	,	.	?
a	0	5	9	6	1	10	4	8	18	9	10	0	0
as	5	4	2	1	0	0	7	10	3	2	1	0	5
chuck	9	2	0	8	0	5	1	9	11	2	4	3	3
could	6	1	8	0	0	4	0	6	8	0	2	2	2
how	1	0	0	0	0	0	4	3	0	2	0	0	0
if	10	0	5	4	0	0	0	0	10	3	8	0	0
much	4	7	1	0	4	0	0	10	2	3	0	0	3
wood	8	10	9	6	3	0	10	2	8	5	0	4	6
woodch.	18	3	11	8	0	10	2	8	0	8	10	1	1
would	9	2	2	0	2	3	3	5	8	0	5	0	0
,	10	1	4	2	0	8	0	0	10	5	0	0	0
.	0	0	3	2	0	0	0	4	1	0	0	0	0
?	0	5	3	2	0	0	3	6	1	0	0	0	0

COALS-STEP 2

Step 2 of the COALS method: Raw counts are converted to correlations.

	a	as	chuck	could	how	if	much	wood	woodch.	would	,	.	?
a	-0.167	-0.014	0.014	0.009	-0.017	0.085	-0.018	-0.033	0.096	0.069	0.085	-0.055	-0.079
as	-0.014	0.031	-0.048	-0.049	-0.037	-0.077	0.133	0.103	-0.054	-0.021	-0.050	-0.037	0.133
chuck	0.014	-0.048	-0.113	0.094	-0.045	0.021	-0.061	0.031	0.048	-0.046	-0.002	0.088	0.031
could	0.009	-0.049	0.094	-0.075	-0.037	0.033	-0.070	0.022	0.049	-0.075	-0.021	0.069	0.023
how	-0.017	-0.037	-0.045	-0.037	-0.018	-0.037	0.192	0.070	-0.055	0.069	-0.037	-0.018	-0.026
if	0.085	-0.077	0.021	0.033	-0.037	-0.077	-0.071	-0.106	0.085	0.006	0.138	-0.037	-0.053
much	-0.018	0.133	-0.061	-0.070	0.192	-0.071	-0.065	0.128	-0.061	0.019	-0.071	-0.034	0.072
wood	-0.033	0.103	0.031	0.022	0.070	-0.106	0.128	-0.113	-0.033	0.001	-0.106	0.111	0.100
woodch.	0.096	-0.054	0.048	0.049	-0.055	0.085	-0.061	-0.033	-0.167	0.049	0.085	-0.017	-0.051
would	0.069	-0.021	-0.046	-0.075	0.069	0.006	0.019	0.001	0.049	-0.075	0.060	-0.037	-0.053
,	0.085	-0.050	-0.002	-0.021	-0.037	0.138	-0.071	-0.106	0.085	0.060	-0.077	-0.037	-0.053
.	-0.055	-0.037	0.088	0.069	-0.018	-0.037	-0.034	0.111	-0.017	-0.037	-0.037	-0.018	-0.026
?	-0.079	0.133	0.031	0.023	-0.026	-0.053	0.072	0.100	-0.051	-0.053	-0.053	-0.026	-0.037

$$r = \frac{T w_{a,b} - \sum_j w_{a,j} \sum_i w_{b,i}}{\sqrt{\sum_j w_{a,j} (T - \sum_j w_{a,j}) \sum_i w_{b,i} (T - \sum_i w_{b,i})}}$$

$$\text{where } T = \sum_i \sum_j w_{i,j}$$

COALS - STEP 3

Step 3 of the COALS method: Negative values discarded and the positive values square rooted.

	a	as	chuck	could	how	if	much	wood	woodch.	would	,	.	?
a	0	0	0.120	0.093	0	0.291	0	0	0.310	0.262	0.291	0	0
as	0	0.175	0	0	0	0	0.364	0.320	0	0	0	0	0.365
chuck	0.120	0	0	0.306	0	0.146	0	0.177	0.220	0	0	0.297	0.175
could	0.093	0	0.306	0	0	0.182	0	0.149	0.221	0	0	0.263	0.151
how	0	0	0	0	0	0	0.438	0.265	0	0.263	0	0	0
if	0.291	0	0.146	0.182	0	0	0	0	0.291	0.076	0.372	0	0
much	0	0.364	0	0	0.438	0	0	0.358	0	0.136	0	0	0.268
wood	0	0.320	0.177	0.149	0.265	0	0.358	0	0	0.034	0	0.333	0.317
woodch.	0.310	0	0.220	0.221	0	0.291	0	0	0	0.221	0.291	0	0
would	0.262	0	0	0	0.263	0.076	0.136	0.034	0.221	0	0.246	0	0
,	0.291	0	0	0	0	0.372	0	0	0.291	0.246	0	0	0
.	0	0	0.297	0.263	0	0	0	0.333	0	0	0	0	0
?	0	0.365	0.175	0.151	0	0	0.268	0.317	0	0	0	0	0

COALS RESULTS

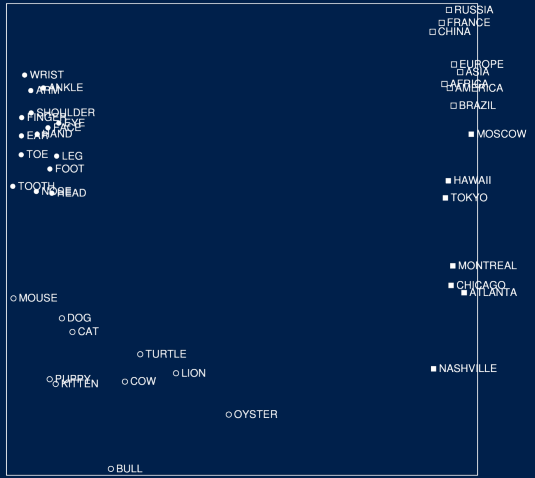
Nearest neighbors and their percent correlation similarities for a set of nouns

	gun		point		mind		monopoly
1)	46.4	handgun	32.4	points	33.5	minds	39.9	monopolies
2)	41.1	firearms	29.2	argument	24.9	consciousness	27.8	monopolistic
3)	41.0	firearm	25.4	question	23.2	thoughts	26.5	corporations
4)	35.3	handguns	22.3	arguments	22.4	senses	25.0	government
5)	35.0	guns	21.5	idea	22.2	subconscious	23.2	ownership
6)	32.7	pistol	20.1	assertion	20.8	thinking	22.2	property
7)	26.3	weapon	19.5	premise	20.6	perception	22.2	capitalism
8)	24.4	rifles	19.3	moot	20.4	emotions	21.8	capitalist
9)	24.2	shotgun	18.9	distinction	20.1	brain	21.6	authority
10)	23.6	weapons	18.7	statement	19.9	psyche	21.3	subsidies

COALS RESULTS - VERBS

	need		buy		play		change	
1)	50.4	want	53.5	buying	63.5	playing	56.9	changing
2)	50.2	needed	52.5	sell	55.5	played	55.3	changes
3)	42.1	needing	49.1	bought	47.6	plays	48.9	changed
4)	41.2	needs	41.8	purchase	37.2	players	32.2	adjust
5)	41.1	can	40.3	purchased	35.4	player	30.2	affect
6)	39.5	able	39.7	selling	33.8	game	29.5	modify
7)	36.3	try	38.2	sells	32.3	games	28.3	different
8)	35.4	should	36.3	buys	29.0	listen	27.1	alter
9)	35.3	do	34.0	sale	26.8	playable	25.6	shift
10)	34.7	necessary	31.5	cheap	25.0	beat	25.1	altering

MULTIDIMENSIONAL SCALING



SOME INSIGHTS 1/2

- The majority of the correlations are negative
- Words with negative correlations do not contribute well to finding similarity than the ones with positive correlation
- Closed-class words (147) convey syntactic information than semantic - could be removed from the correlation table punctuation marks, she, he, where, after, ...

SOME INSIGHTS 2/2

- **Positive Correlation** means that two words often appear together. In other words, their contexts are similar.
  - CMI and Prodigy have a strong correlation
- **Zero correlation** means the pair of words are statistically independent. Hence no influence
  - CMI and Engineering_Drawing/blueprint have no inherent or direct relationship in terms or context.
- **Negative correlation** indicates an inverse relationship. For every word pair in this set, the second word in each pair is less likely to appear, and vice versa.
  - When talking about one of the specializations of CMI as the first word in a pair, some words, such as art, literature, philosophy, and religion, are less likely to appear together (although they may appear a few times).

FIRST AND SECOND ORDER ASSOCIATIONS OF DSM

First Order Association

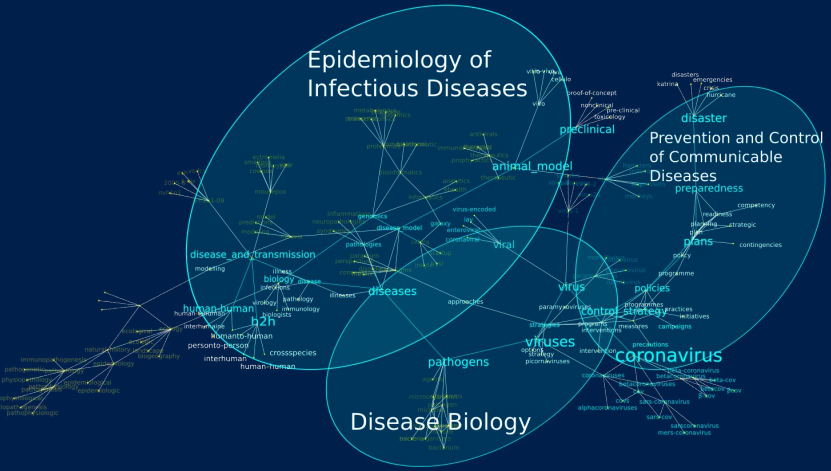
- Pairs of words in common contexts are semantically related
- If a word w_x occurred in several contexts along with w_y, then w_x and w_y are related by the first-order association. w_x and w_y are called as first-order associates.
- For any pair of words, w_i and w_j, if the strength of similarity is stronger, then they have a large number of common first-order associates

Second Order Association

- If a word w_y occurred in several contexts along with w_z in which w_x is absent (or occurred in a statistically insignificant number of times), then w_x and w_z are related by the second-order association. w_x and w_z are called as second-order associates.

  $\forall w_i, w_j, w_k \in V$, if we have $w_x R w_y$ and $w_y R w_z$, then $w_x R w_z$ where the relation R is transitive

MULTI-LEVEL WORD ASSOCIATIONS



DENSE VECTORS

- Not sparse
- Shorter than sparse word vectors
- Real-valued and continuous
- Captures fine-grained semantic information
- Can be used to represent the entire sentences or paragraphs
- Word2Vec, GloVe, and FastText use dense embeddings
- Typical sizes are 50, 100, or 300 elements

REFERENCES

[1] Douglas LT Rohde, Laura M Gonnerman, and David C Plaut. *An improved model of semantic similarity based on lexical co-occurrence*. 2006. URL: https://cnbc.cmu.edu/~plaut/papers/pdf/RohdeGonnermanPlautSUB-CogSci.COALS.pdf.

[2] Omer Levy, Yoav Goldberg, and Ido Dagan. "Improving distributional similarity with lessons learned from word embeddings". In: *Transactions of the association for computational linguistics* 3 (2015), pp. 211–225.

[3] Will Lowe. "Towards a theory of semantic space". In: *Proceedings of the annual meeting of the cognitive science society*. Vol. 23. 23. 2001.

[4] Sebastian Padó and Mirella Lapata. "Dependency-Based Construction of Semantic Space Models". In: *Computational Linguistics* 33.2 (2007), pp. 161–199. DOI: 10.1162/coli.2007.33.2.161. URL: https://aclanthology.org/J07-2002.