ACL 2019

Towards Understanding Linear Word Analogies

Kawin Ethayarajh, David Duvenaud, Graeme Hirst

MOTIVATION

Why can vector arithmetic be used to operate on word embeddings generated by non-linear models?

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

Current theories make untenable assumptions about the word frequency distribution or embedding space.

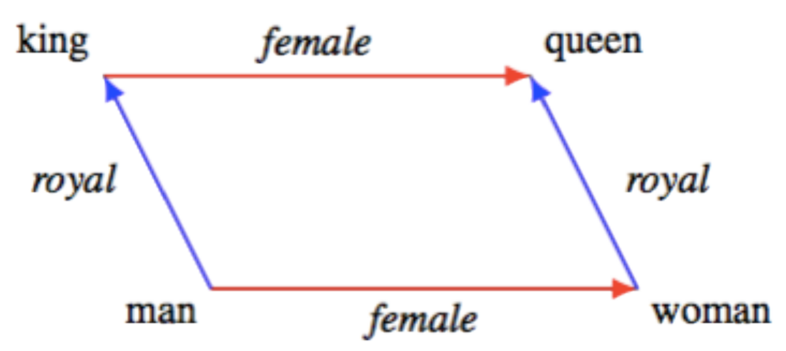
THE STRUCTURE OF WORD ANALOGIES

Definitions

A **word analogy** is an invertible transformation f that holds over a set of ordered word pairs iff

$$\forall\ (x, y) \in S, f(x) = y \wedge f^{-1}(y) = x$$

When f is of the form $\vec{x} \mapsto \vec{x} + \vec{r}$, it is a **linear word analogy.** When linear word analogies hold exactly, they form a parallelogram in the embedding space:



Interpreting Inner Products

GloVe and SGNS implicitly factorize a word-context matrix containing a co-occurrence statistic (Levy and Goldberg, 2014).

- f holds over S in an SGNS or GloVe word space iff $g : \vec{x}_c \mapsto \vec{x}_c + \lambda\vec{r}$ holds in the corresponding context space.
- We can write $\|\vec{x} - \vec{y}\|^2$ as the inner product $\langle \vec{x} - \vec{y}, \vec{x}_c - \vec{y}_c \rangle$ scaled by $1/\lambda$.

THEORETICAL RESULTS

What conditions have to be satisfied by the training corpus for these linear word analogies to hold in a noiseless space?

Co-occurrence Shifted PMI Theorem

Let the co-occurrence shifted PMI be **csPMI**(x, y) = **PMI**$(x, y) + \log p(x, y)$, W be a noiseless SGNS or GloVe word space, M be the word-context matrix that is implicitly factorized, and S a set of ordered word pairs.

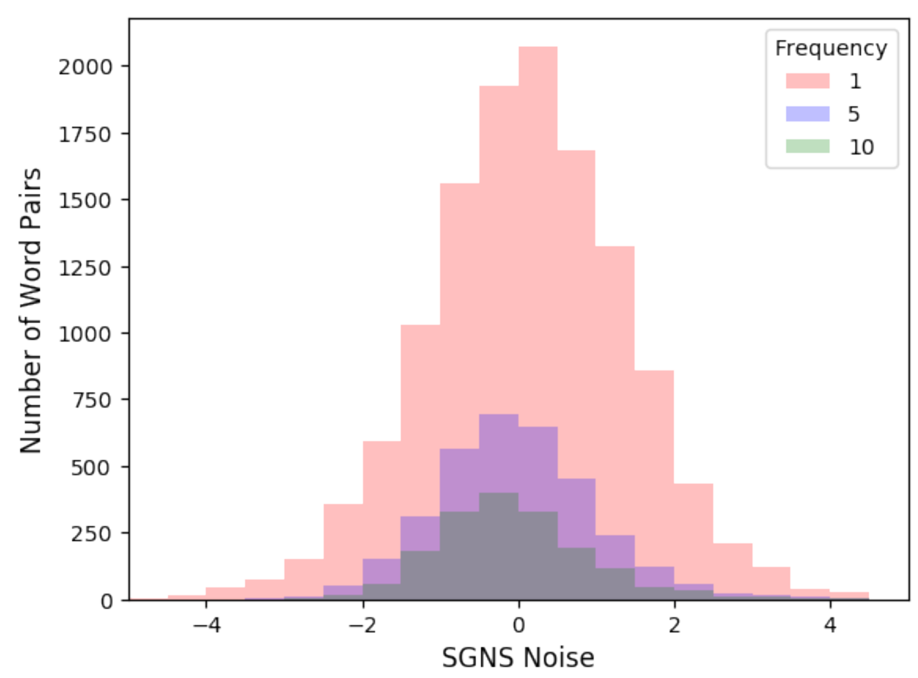
A linear analogy f holds over S iff

- csPMI(x, y) is the same for every word pair in S
- csPMI(x, y) = csPMI(a, b) for any two word pairs in S
- $\{M_{a,\cdot} - M_{y,\cdot}, M_{b,\cdot} - M_{y,\cdot}, M_{x,\cdot} - M_{y,\cdot}\}$ is linearly dependent ("contextually coplanar")

Robustness to Noise

In practice, word analogies are quite robust to noise. Why?

- The definition of vector equality is looser in practice: $(a, ?) : (x, y)$ is solved by finding the word vector *closest* to $\vec{a} + (\vec{y} - \vec{x})$.
- Analogies mostly hold over frequent word pairs, which are associated with less noise.



COROLLARIES & EMPIRICAL RESULTS

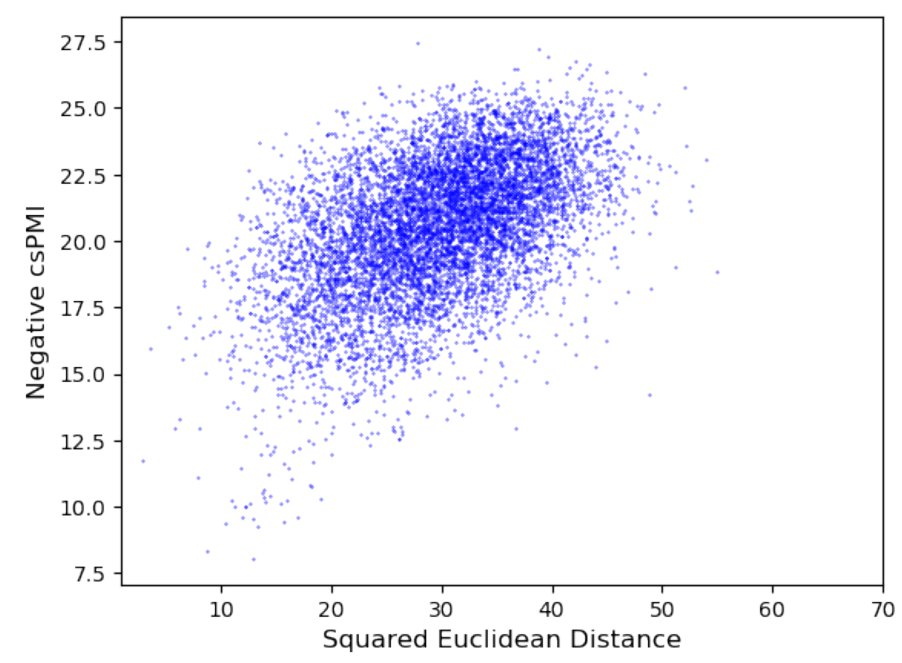
1. We prove the long-standing conjecture (Pennington et al., 2014) that "*a is to b as x is to y*" holds iff for every word w,

$$\frac{p(w\,|\,a)}{p(w\,|\,b)} \approx \frac{p(w\,|\,x)}{p(w\,|\,y)}$$

2. In a noiseless space, the squared Euclidean distance between words is a decreasing linear function of csPMI:

$$\lambda\|\vec{x} - \vec{y}\|^2 = -\text{csPMI}(x, y) + \alpha$$

Empirically, the correlation is quite strong (Pearson's $r = 0.502$):



3. The change in mean csPMI mirrors a change in the type of analogy, from geography to verb tense to adjectives:

Analogy	Mean csPMI	Mean PMI
capital-world	−9.294	6.103
capital-common-countries	−9.818	4.339
city-in-state	−10.127	4.003
gram6-nationality-adjective	−10.691	3.733
family	−11.163	4.111
gram8-plural	−11.787	4.208
gram5-present-participle	−14.530	2.416
gram9-plural-verbs	−14.688	2.409
gram7-past-tense	−14.840	1.006
gram3-comparative	−15.111	1.894
gram2-opposite	−15.630	2.897
gram4-superlative	−15.632	2.015
currency	−15.900	3.025
gram1-adjective-to-adverb	−17.497	1.113

4. When the variance in csPMI is lower, analogy solutions are more accurate (Pearson's $r = -0.70$).

REFERENCES

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In Proceedings of EMNLP, 1532–1543

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In Advances in Neural Information Processing Systems, 2177–2185

Acknowledgements: Thanks to Omer Levy and Yoav Goldberg for their comments.

UNIVERSITY OF TORONTO