NLP and Word Embeddings

deeplearning.ai

GloVe word vectors

This is not used as much as the Word2Vec or the skip-gram models, but it has some enthusiasts, because, I think, in part of its simplicity.

GloVe (global vectors for word representation)

I want a glass of orange juice to go along with my cereal.

c, t

X_{ij} = # times i appears in context of j.

$\underset{c\ \ t}{\uparrow\ \uparrow}$ $\qquad$ $\underset{t}{\uparrow}$ (i) $\qquad$ $\underset{c}{\uparrow}$ (j)

$X_{ij} = X_{ji}$

In fact, if you defining context and target in terms of whether or not they appear plus minus 10 words of each other, then it would be a symmetric relationship.
Although, if your choice of context was that context is always the word immediately before the target word, then x_ij and x_ji may not be symmetric like this.
So x_ij is a count that captures how often do words i and j appear with each other or close to each other.

[Pennington et. al., 2014. GloVe: Global vectors for word representation]

Model

$$\text{minimize} \sum_{i=1}^{10,000} \sum_{j=1}^{10,000} \underbrace{f(X_{ij})}_{\text{weighting term}} \left(\theta_i^T e_j + b_i + b_j' - \log X_{ij} \right)^2$$

$t \quad c$

"$\theta_t^T e_c$"

0?

$f(X_{ij}) = 0$ of $X_{ij} = 0$. "$0 \log 0$" $= 0$

this, is, of, a, ...

durian

θ_i, e_j are symmetric

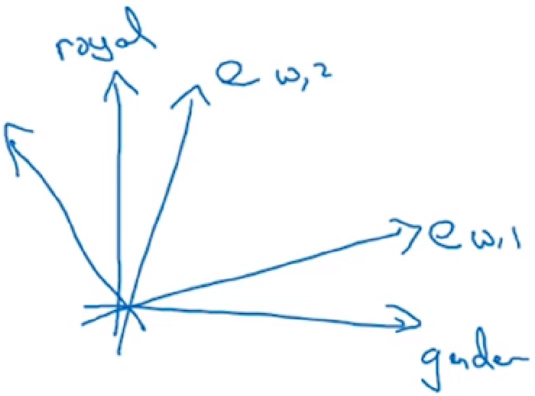
$$e_w^{(final)} = \frac{e_w + \theta_w}{2}$$

The algorithm was building on the history of much more complicated algorithms

Andrew Ng

A note on the featurization view of word embeddings

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)
Gender	−1	1	-0.95	0.97
Royal	0.01	0.02	0.93	0.95
Age	0.03	0.02	0.70	0.69
Food	0.09	0.01	0.02	0.01



$$\text{minimize} \sum_{i=1}^{10,000} \sum_{j=1}^{10,000} f(X_{ij}) \left(\theta_i^T e_j + b_i - b_j' - \log X_{ij} \right)^2$$

$$(A\theta_i)^T (A^{-T} e_j) = \theta_i^T \cancel{A^T} \cancel{A^{-T}} e_j$$

Andrew Ng

Don't worry if you do not follow the linear algebra, but that's a brief proof that shows that with an algorithm like this, you can't guarantee that the axis used to represent the features will be well-aligned with that might be easily humanly interpretable axis. In particular, the first feature might be a combination of gender, and royal, and age, and food, and cost, and size, it is a noun or an action verb, all the other features. So it's very difficult to look at individual components, individual rows of the embedding matrix and assign a human interpretation to that. But despite this type of linear transformation, the parallelogram map that we worked out when we were describing analogies, that still works. (potentially arbitrary linear transformation)