# A latent variable approach to word embeddings

Sanjoy Dasgupta

## An embedding-based generative model of text

S. Arora, Y. Li, Y. Liang, T. Ma and A. Risteski. Rand-Walk: a latent variable model approach to word embeddings, 2015.

Common methodology in unsupervised learning:

- Define a generative model, with some parameters $\theta$, that produces a particular kind of data.
- Think of observed data as coming from such a model.
- Option 1: find the $\theta$ most likely to have produced the data.
- Option 2: assume the data truly came from such a model and recover the "true" $\theta$.

# Overview of generative model

Want a generative model of text under which:

- Word embedding methods can be seen to recover the "correct" vectors.
- Certain types of relations can be seen to indeed satisfy simple linear relationships in vector space.

Generative process for corpus of words: $w_1 w_2 w_3 \cdots$:

- Unseen "discourse vector" $c_t$ follows a random walk in $\mathbb{R}^d$:

$$c_1 \rightarrow c_2 \rightarrow c_3 \rightarrow \cdots$$

- At time $t$, word $w \in W$ is emitted, with probability

$$\Pr(w) \propto e^{v(w) \cdot c_t}.$$

Here $\{v(w) : w \in W\} \subset \mathbb{R}^d$ are embeddings to be recovered.

# Assumptions on discourse vector

- Stationary distribution of $c_t$ is uniform over the unit sphere in $\mathbb{R}^d$.
- Each step of the discourse vector is small:

$$\|c_t - c_{t-1}\| \leq \epsilon/\sqrt{d}.$$

# Assumptions on word vectors

Each word vector $v(w)$ is chosen independently as follows:

- Pick a direction $v$ at random from the unit sphere in $\mathbb{R}^d$
- Pick a length $s$ with mean $O(1)$ and max value some small constant.
- Word vector $= sv$.

Why bother with $s$?

When discourse vector is $c$,

$$\text{Prob}_c(w) = \frac{e^{c \cdot v(w)}}{Z_c}, \quad \text{where} \quad Z_c = \sum_{w'} e^{c \cdot v(w')}$$

**Claim.** The normalizers $Z_c$ are almost identical for almost all $c$.

# Pointwise mutual information

Under these assumptions, for any $w, w'$,

$$\text{PMI}(w, w') \;=\; \log \frac{\Pr(w, w')}{\Pr(w)\Pr(w')} \;\approx\; \frac{v(w) \cdot v(w')}{d}$$

Therefore, can recover $v(\cdot)$ by factoring the PMI matrix.

This is shown via two intermediate results:

$$\log \Pr(w) \approx \frac{\|v(w)\|^2}{2d} - \log Z$$
$$\log \Pr(w, w') \approx \frac{\|v(w) + v(w')\|^2}{2d} - 2 \log Z$$

# Probability of a word

**Claim.** $\log \Pr(w) \approx \frac{\|v(w)\|^2}{2d} - \log Z$.

Since discourse vector $c$ is uniform over $S^{d-1}$,

$$\begin{aligned}
\Pr(w) &= \mathbb{E}_{c \sim S^{d-1}} \left[ \frac{e^{c \cdot v(w)}}{Z_c} \right] \\
&\approx \frac{1}{Z} \mathbb{E}_{c \sim N(0,(1/d)I_d)} [\exp(c \cdot v(w))] \\
&\approx \frac{1}{Z} \mathbb{E}_{x \sim N(0,\|v(w)\|^2/d)} [\exp(x)] \\
&= \frac{1}{Z} \exp(\|v(w)\|^2/(2d))
\end{aligned}$$

# Why are relations = lines?

Why can we solve analogies $a : b :: c :?$ using

$$\arg\min_d \|v(b) - v(a) + v(c) - v(d)\|^2 \ ?$$

Previous insight (Pennington, Socher, Manning): For any word $x$,

$$\frac{\Pr(x|\text{king})}{\Pr(x|\text{queen})} = \frac{\Pr(x|\text{man})}{\Pr(x|\text{woman})}$$

To see this, consider three cases:

- $x$ is gender-neutral
- $x$ is 'he', 'Henry', etc
- $x$ is 'she', 'Elizabeth', etc.

# Relations, cont'd

Key assumption: There is a function $\Phi : W \to \mathbb{R}$ such that for any pair $(a, b)$ satisfying the relation,

$$\frac{\Pr(x|a)}{\Pr(x|b)} \approx \Phi(x).$$

At the same time, under earlier assumptions,

$$\log \Pr(x|a) \;=\; \mathrm{PMI}(x, a) + \log \Pr(x) \;\approx\; \frac{v(x) \cdot v(a)}{d} + \log \Pr(x)$$

and thus

$$\log \frac{\Pr(x|a)}{\Pr(x|b)} \;\approx\; v(x) \cdot \frac{v(a) - v(b)}{d}.$$

Therefore, $v(a) - v(b)$ is the same for all $(a, b)$ satisfying the relation.