词嵌入表示

N-gram Model Recap

N-元模型回顾

- 句子: $x_1, \dots x_m$
- 贝叶斯展开:

$$P(x_1, x, \dots, x_m) = P(x_1)P(x_2|x_1) \dots P(x_m|x_1, x_2, \dots, x_{m-1})$$

- N-元模型:

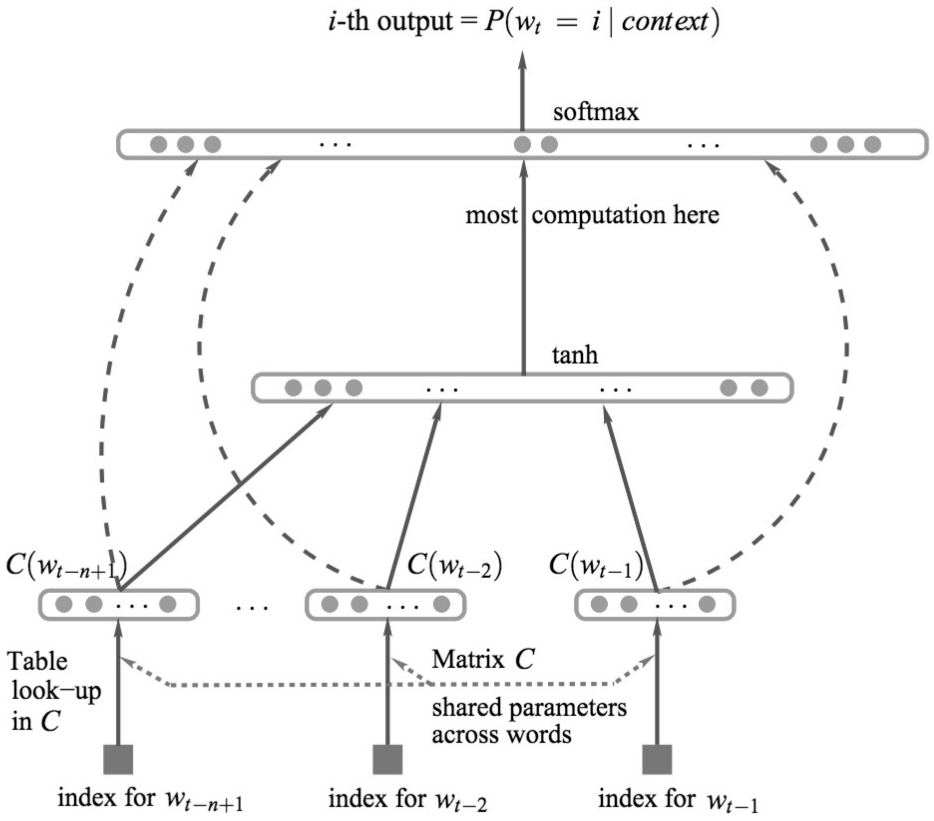
$$P(x_1, x_2, \dots, x_m) = P(x_1)P(x_2|x_1) \dots P(x_m|x_1, x_2, \dots, x_{m-1})$$

- 极大似然估计:

$$P\left(x_i \middle| x_{i-(n-1)}, \dots, x_{i-1}\right) = \frac{count(x_{i-(n-1)}, \dots, x_{i-1}, x_i)}{count(x_{i-(n-1)}, \dots, x_{i-1})}$$

Neural Language Model

神经语言模型



A Neural Probabilistic Language Model, Yoshua Bengio, etc

Neural Language Model

神经语言模型

- “一只狗在叫 ”出现一百次   v.s.   “一只猫在叫”出现一次
- 对n-gram模型来说, 狗和猫的权重不一致
- 对神经概率语言模型来说:
    - 假定了相似词对于的词向量也相似
    - 概率函数关于词向量是光滑的

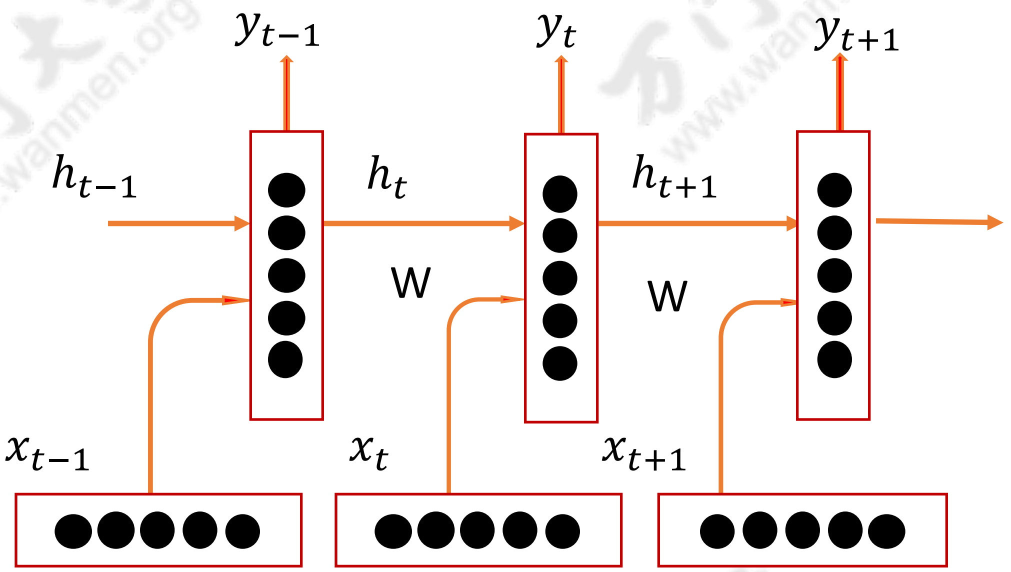
RNN Language Model

递归神经网络语言模型

- 句子: $x_1, \dots x_m$
- 时间t时刻:

$$h_t = \sigma\left(W^{(hh)} h_{t-1} + W^{(hx)} e(x_t)\right)$$
$$\hat{y}_t = softmax(W^{(s)} h_t)$$
$$\hat{P}\left(y_{t+1} = v_j \middle| x_t, \dots, x_1\right) = \hat{y}_{t,i}$$

- 结构



Word Embedding

词嵌入

- 独热表示(One-hot Encoding):

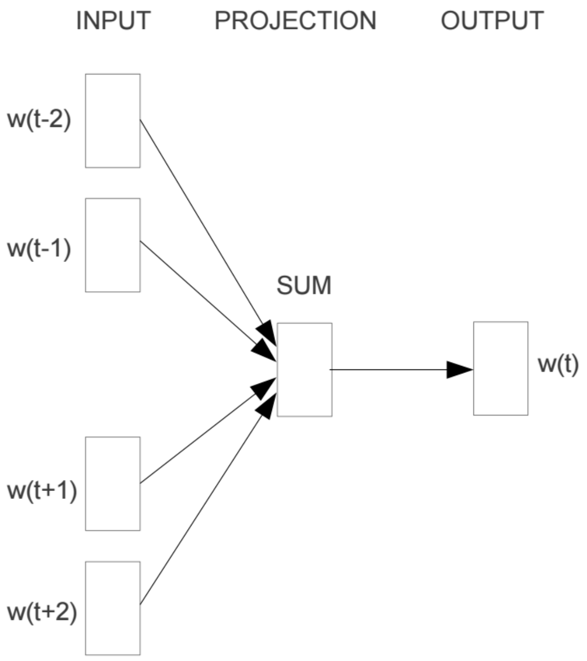
国王	0	1	0	0	0
王后	0	0	0	1	0

- 分布式表示(Distributed Representation):

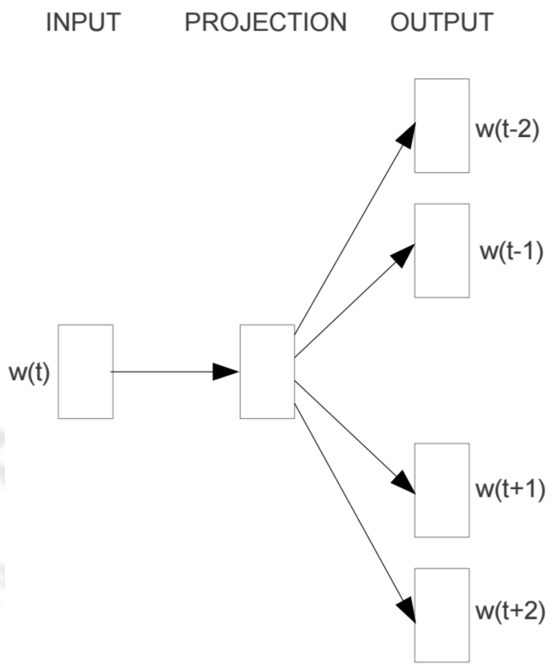
国王	4.3	1.0	0.6	3.2	-0.7
王后	0.25	0.5	-1.2	0.9	1.2

Word Embedding: Skip-Gram & CBOW

词嵌入: Skip-Gram & CBOW



CBOW



Skip-gram

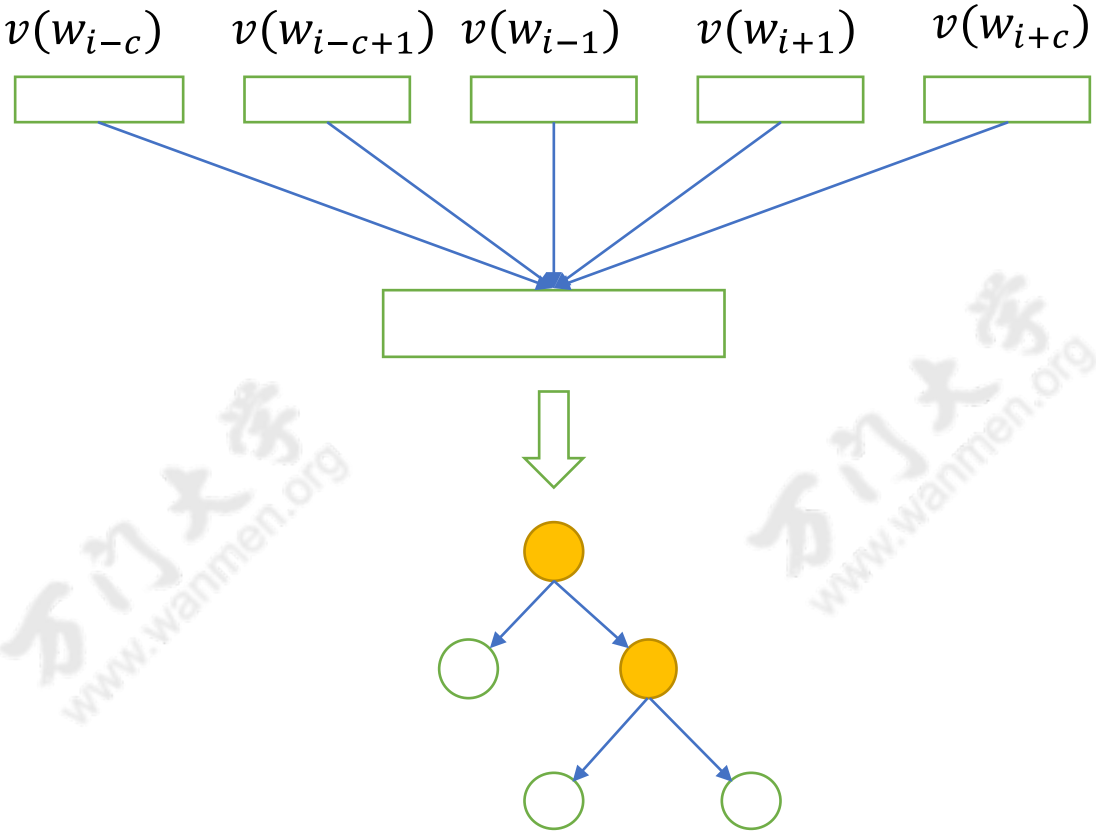
CBOW: Hierarchical Softmax

连续词袋模型: 分层Softmax

输入层

投影层

输出层



$$X_w = \sum_{\substack{j=i-c \\ j\neq i}}^{i+c} v(w_j)$$

Huffman树

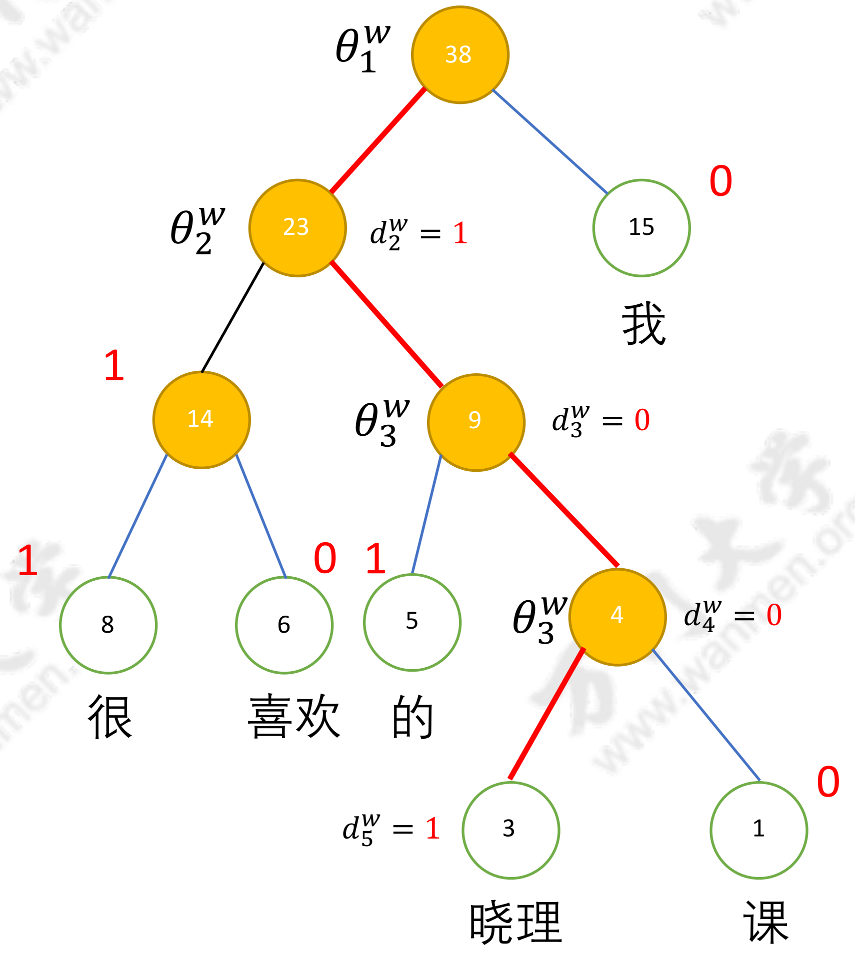
CBOW: Hierarchical Softmax

连续词袋模型: 分层Softmax

- p^w: 从根结点出发达到w对应的叶子结点的路径
- n^w: 路径p^w中包含的结点的个数
- $p_1^w, p_2^w, \ldots, p_{n^w}^w$: 路径$p^w$中的$n^w$结点, p_1^w为根结点, $p_{n^w}^w$为词w对应的结点
- $d_2^w, d_3^w, \ldots, p_{n^w}^w \in \{0,1\}$: 词$w$的Huffman编码, 由$l^w - 1$位编码构成, d_j^w表示路径p^w中第j个结点对应的编码
- $\theta_1^w, \theta_2^w, \ldots, \theta_{n^w}^w \in R^m$: 路径$p^w$中非叶子结点对应的向量

CBOW: Hierarchical Softmax

连续词袋模型: 分层Softmax



- w='晓理'
- $n^w = 5$
- $d_2^w, d_3^w, d_4^w, d_5^w$ 为1,0,0,1, '晓理'的Huffman编码
- 每一次分支都视为一次二分类
- $Label(p_i^w) = 1 - d_i^w, i = 2,3, \dots, l^w$
- 分为正类概率: $\sigma(X_w^T\theta) = \frac{1}{1+e^{-X_w^T\theta}}$

CBOW: Hierarchical Softmax

连续词袋模型: 分层Softmax

- 计算$p(w_i|w_{i-c},\dots,w_{i-1},w_{i+1},\dots,w_{i+c})$
- 第i次分支概率: $p(d_i^w|X_w,\theta_{i-1}^w)=\sigma(X_w^T\theta_{i-1}^w)^{1-d_i^w}(1-\sigma(X_w^T\theta_{i-1}^w))^{d_i^w}$
- $p(w_i|w_{i-c},\dots,w_{i-1},w_{i+1},\dots,w_{i+c})=\prod_{j=2}^{n^w}p(\,d_i^w|X_w,\theta_{j-1}^w)$
- 例如: $p(\text{晓理}|w_{i-c},\dots,w_{i-1},w_{i+1},\dots,w_{i+c})=\prod_{j=2}^{5}p\,(d_i^w|X_w,\theta_{j-1}^w)$
- 目标函数: $\mathcal{L}=\sum_{w\in C}\log(p(w|\,w_{i-c},\dots,w_{i-1},w_{i+1},\dots,w_{i+c})$

CBOW: Hierarchical Softmax

连续词袋模型: 分层Softmax

$$\mathcal{L} = \sum_{w \in C} log \prod_{j=2}^{n^w} \left\{ \sigma\left(X_w^T \theta_{j-1}^w\right)^{1-d_j^w} \left(1 - \sigma\left(X_w^T \theta_{j-1}^w\right)\right)^{d_j^w} \right\}$$

$$\mathcal{L}(w,j) = \left(1 - d_j^w\right) \cdot \log\left(\sigma\left(X_w^T \theta_{j-1}^w\right)\right) + d_j^w \cdot \log(1 - \sigma\left(X_w^T \theta_{j-1}^w\right))$$

$$\frac{\partial \mathcal{L}(w,j)}{\partial \theta_{j-1}^w} = \left[1 - d_j^w - \sigma\left(X_w^T \theta_{j-1}^w\right)\right] X_w$$

$$\frac{\partial \mathcal{L}(w,j)}{\partial X_w} = \left[1 - d_j^w - \sigma\left(X_w^T \theta_{j-1}^w\right)\right] \theta_{j-1}^w$$

$$v(\tilde{w}) = v(\tilde{w}) + \lambda \sum_{j=2}^{n^w} \frac{\partial \mathcal{L}(w,j)}{\partial X_w}, \qquad \tilde{w} \in Context(w)$$

Skip-gram: Hierarchical Softmax

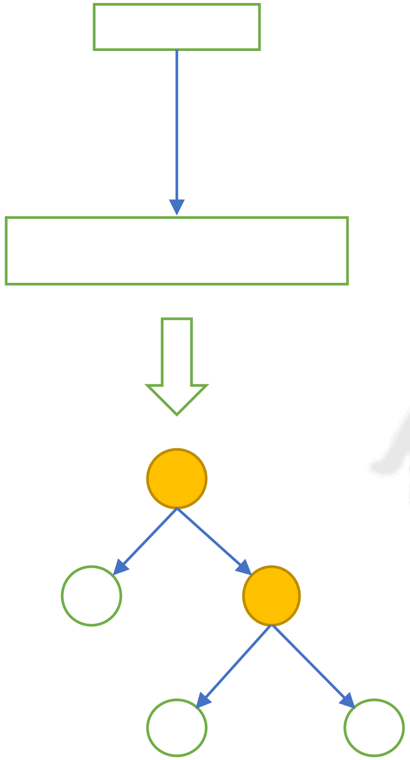
Skip-gram: 分层Softmax

$v(w_i)$

输入层

投影层

$v(w_i)$



输出层

Huffman树

Skip-gram: Hierarchical Softmax

Skip-gram: 分层Softmax

- 计算$p(w_{i-c},\dots,w_{i-1},w_{i+1},\dots,w_{i+c}|w_i)$
- 第j次分支概率: $p\left(d_j^u \middle| v(w),\theta_{j-1}^u\right)=\sigma\left(v(w)^T\theta_{j-1}^u\right)^{1-d_j^u}\left(1-\sigma\left(v(w)^T\theta_{j-1}^u\right)\right)^{d_j^u}$
- $p(w_{i-c},\dots,w_{i-1},w_{i+1},\dots,w_{i+c}|w_i)=\prod_{j=2}^{n^u}p(\,d_i^u|v(w),\theta_{j-1}^u)$
- 目标函数: $\mathcal{L}=\sum_{w\in C}\log\prod_{u\in context(w)}\prod_{j=2}^{n^u}\left\{\sigma\left(v(w)^T\theta_{j-1}^u\right)^{1-d_j^u}\left(1-\sigma\left(v(w)^T\theta_{j-1}^u\right)\right)^{d_j^u}\right\}$

Skip-gram: Hierarchical Softmax

Skip-gram: 分层Softmax

$$\mathcal{L} = \sum_{w \in C} log \prod_{w \in Context(w)} \prod_{j=2}^{n^u} \left\{ \sigma\left(v(w)^T \theta_{j-1}^u\right)^{1-d_j^u} \left(1 - \sigma\left(v(w)^T \theta_{j-1}^u\right)\right)^{d_j^u} \right\}$$

$$\mathcal{L}(w,u,j) = \left(1 - d_j^u\right) \cdot \log\left(\sigma\left(v(w)^T \theta_{j-1}^u\right)\right) + d_j^u \cdot \log\left(1 - \sigma\left(v(w)^T \theta_{j-1}^u\right)\right)$$

$$\frac{\partial \mathcal{L}(w,u,j)}{\partial \theta_{j-1}^u} = \left[1 - d_j^w - \sigma\left(v(w)^T \theta_{j-1}^u\right)\right] v(w)$$

$$\frac{\partial \mathcal{L}(w,u,j)}{\partial v(w)} = \left[1 - d_j^u - \sigma\left(v(w)^T \theta_{j-1}^u\right)\right] \theta_{j-1}^u$$

$$v(w) = v(w) + \sum_{u \in Context(w)} \sum_{j=2}^{l^u} \frac{\partial \mathcal{L}(w,u,j)}{\partial v(w)}, \qquad \tilde{w} \in Context(w)$$

CBOW: Negative Sampling

连续词袋模型: 负采样

- 给定w的上下文, 预测w, w为正样本, 其余词为负样本
- $L^w(\widetilde{w}) = \begin{cases} 1, \widetilde{w} = w \\ 0, \widetilde{w} \neq w \end{cases}$
- $p(u|\widetilde{w}) = [\sigma(v(\widetilde{w})^T\theta^u)]^{L^w(u)} \cdot [1-\sigma(v(\widetilde{w})^T\theta^u)]^{1-L^w(u)}$
- 目标函数: $\mathcal{L} = \sum_{w\in C}\log\prod_{\widetilde{w}\in\{w\}\cup NEG^{\widetilde{w}}(w)}\prod_{j=2}^{n^u}\{[\sigma(v(\widetilde{w})^T\theta^u)]^{L^w(u)}[1-\sigma(v(\widetilde{w})^T\theta^u)]^{1-L^w(u)}\}$
- $\frac{\partial\mathcal{L}(w,\widetilde{w},u)}{\partial\theta^u} = [L^w(u)-\sigma(v(\widetilde{w})^T\theta^u]v(\widetilde{w})$
- $\frac{\partial\mathcal{L}(w,\widetilde{w},u)}{\partial v(\widetilde{w})} = [L^w(u)-\sigma(v(\widetilde{w})^T\theta^u]\theta^u$
- $v(\widetilde{w}) = v(\widetilde{w}) + \sum_{u\in\{w\}\cup NEG(w)}\frac{\partial\mathcal{L}(w,u)}{\partial X_w}$, $\widetilde{w}\in Context(w)$

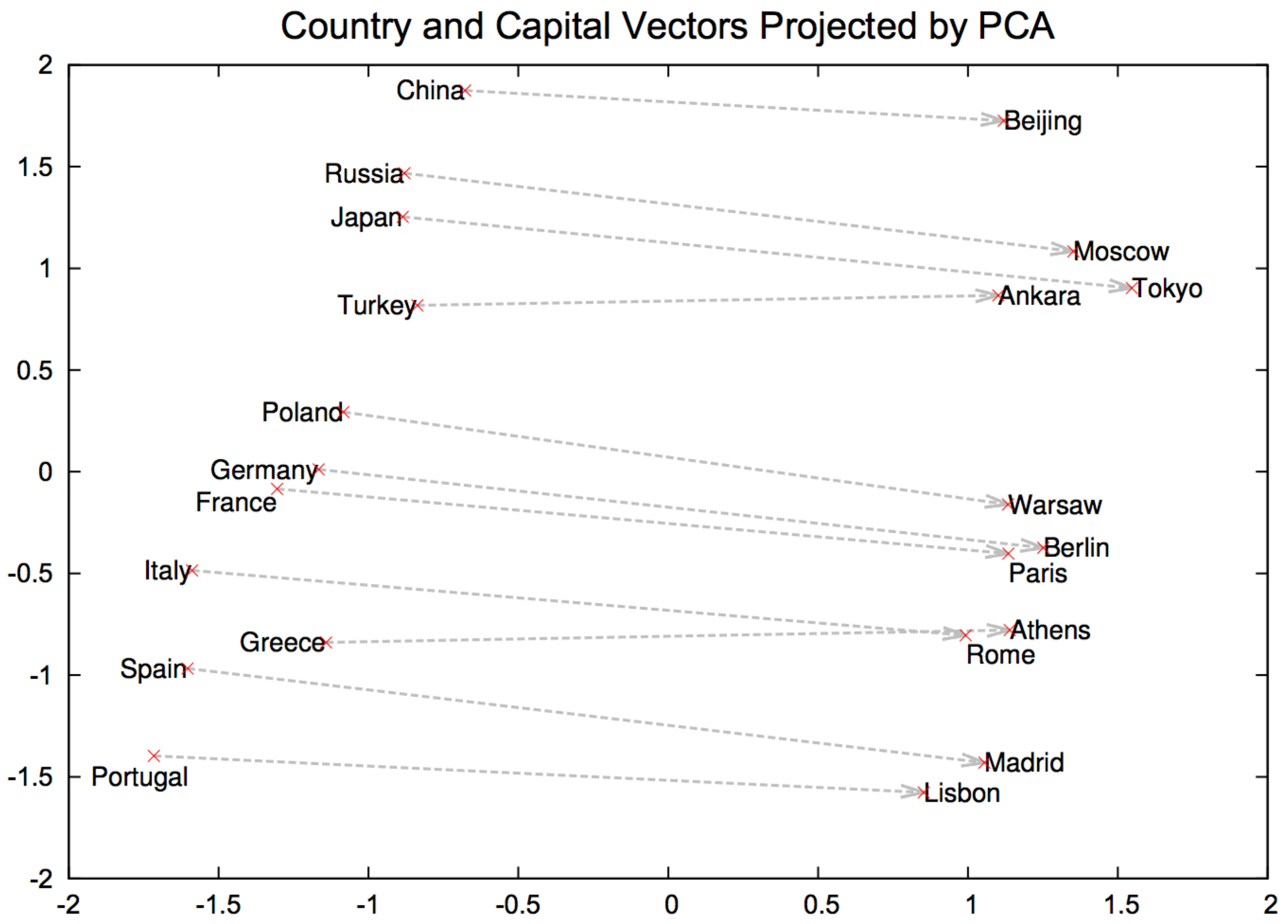
Skip-gram: Negative Sampling

Skip-gram: 负采样

- 给定w的上下文, 预测w, w为正样本, 其余词为负样本
- $L^w(\tilde{w}) = \begin{cases} 1, \tilde{w} = w \\ 0, \tilde{w} \neq w \end{cases}$
- $p(u|Context(w)) = [\sigma(X_w^T\theta^u)]^{L^w(\tilde{w})} \cdot [1 - \sigma(X_w^T\theta^u)]^{\,1-L^w(\tilde{w})}$
- 目标函数: $\mathcal{L} = \sum_{w\in C} \log \prod_{u\in\{w\}\cup NEG(w)} \prod_{j=2}^{n^u} \left\{[\sigma(X_w^T\theta^u)]^{1-d_j^u}[1-\sigma(X_w^T\theta^u)]^{d_j^u}\right\}$
- $\frac{\partial\mathcal{L}(w,u)}{\partial\theta^u} = [L^w(u) - \sigma(X_w^T\theta^u)]X_w$
- $\frac{\partial\mathcal{L}(w,u)}{\partial X_w} = [L^w(u) - \sigma(X_w^T\theta^u]\theta^u$
- $v(\tilde{w}) = v(\tilde{w}) + \sum_{u\in\{w\}\cup NEG(w)} \frac{\partial\mathcal{L}(w,u)}{\partial X_w},\ \tilde{w} \in Context(w)$

Word Vectors: Visualization

词向量: 可视化



Distributed Representations of Words and Phrases and their Compositionality