How to generate a good word embedding

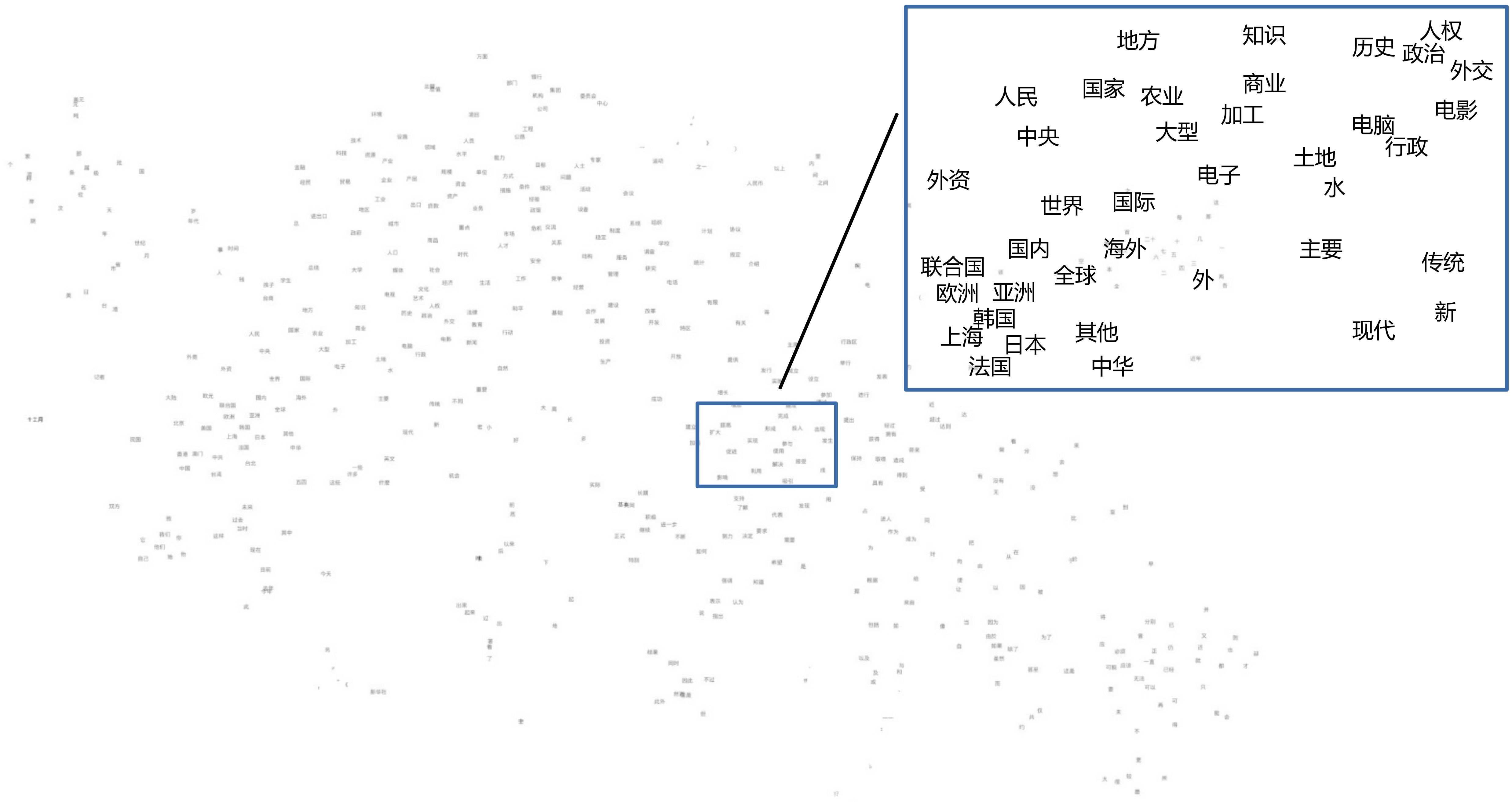
刘 康

中国科学院自动化研究所
模式识别国家重点实验室
2015年8月25日

词表示

- One-hot Word Representation
  - 减肥 [0 0 0 1 0 0 0 0 0 0]
  - 瘦身 [1 0 0 0 0 0 0 0 0 0]
- Distributed Word Representation
  - 减肥 [0.792, –0.177, –0.107, 0.109, –0.542]
  - 瘦身 [0.856, –0.523, 0, 0.2, -0.2]

词表示


地方
知识
历史
人权
政治
外交
人民
国家
农业
商业
加工
电影
电脑
中央
大型
行政
土地
外资
电子
水
世界
国际
国内
海外
主要
联合国
全球
传统
外
欧洲
亚洲
韩国
新
上海
其他
现代
日本
法国
中华

词向量表示的核心

- 利用上下文信息进行词表示
  - 具有相同(类似)上下文信息的词应该具有相同(类似)的词表示[Z. Harris, 1954]

$$\vec{v} = (c_1, c_2, ..., c_n)$$

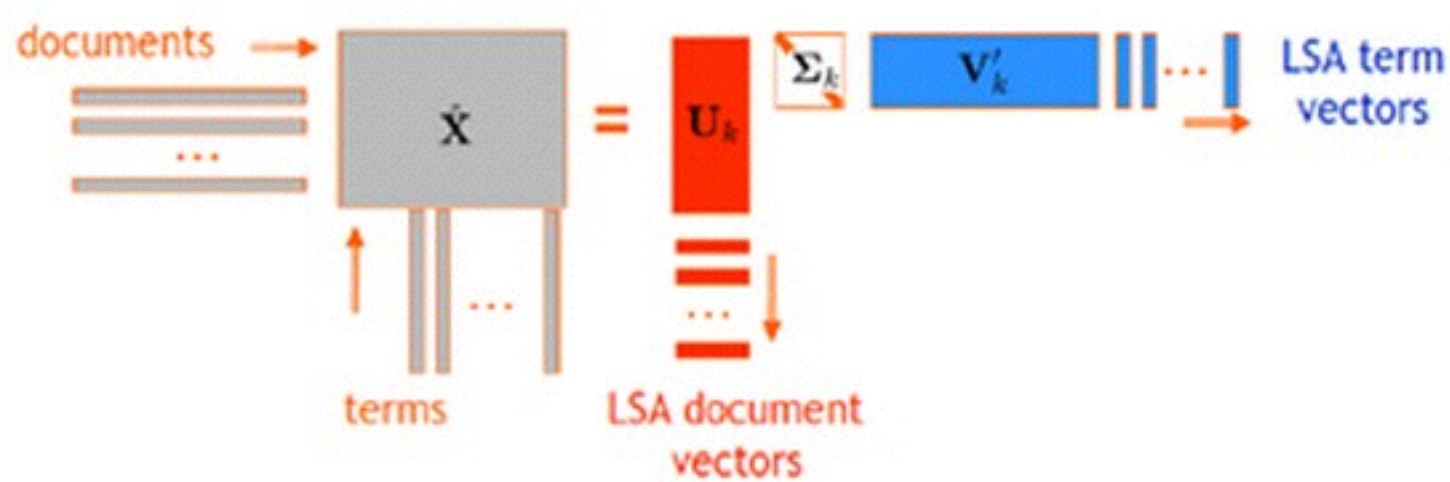
  - 两种上下文选择 [Sahlgren 2006]
    - “词-文档”共现矩阵
    - “词-词”共现矩阵

- Syntagmatic Relation
- Paradigmatic Relation

传统词向量方法

- “词-文档”共现矩阵
  - LSA、PLSA

	d1	d2	d3
w1	1	1	3
w2	2	2	1
w3	4	2	1
w4		3	



$$X \approx U \Sigma V^T$$

传统词向量方法

- “词-文档”矩阵
  - Syntagmatic Relation（组合关系/一阶关系）: Two words are similar if they tend to appear in the contexts of each other
  - Use **co-occurrence events** for building the word space as a syntagmatic use of context [Sahlgren 2006]

I like nature language processing
You like machine learning
We like deep learning

deep→learning
machine→learning

	d1	d2	d3
I	*1*		
like	*1*	*1*	*1*
nature	*1*		
language	*1*		
processing	*1*		
You		*1*	
machine		*1*	
learning		*1*	*1*
We			*1*
deep			*1*

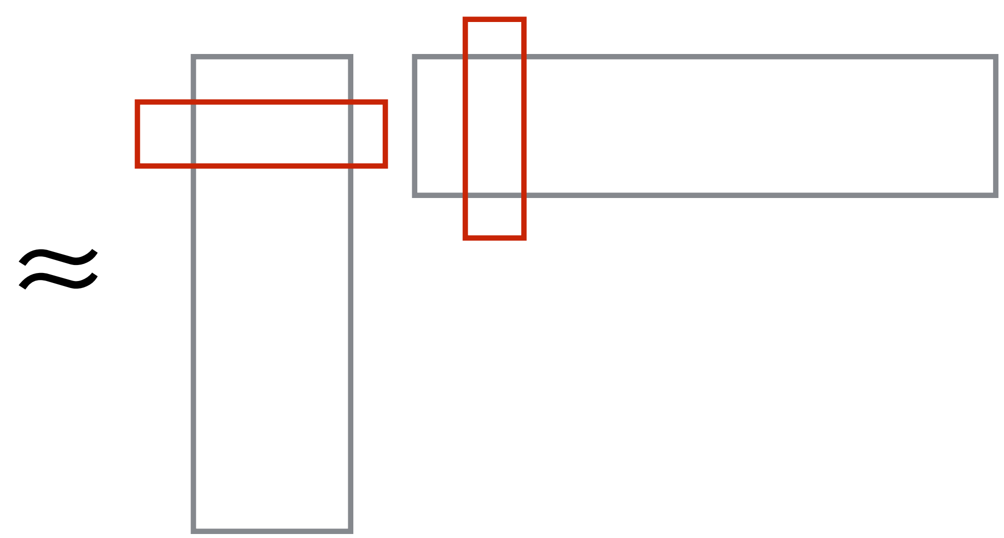
传统词向量方法

- “词-词”共现矩阵
  - HAL [Lund et al. 1996]、GloVe [Pennington et al 2014]

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

词向量　词向量　词词共现

	w1	w2	w3	w4
w1		2	4	1
w2	2		3	
w3	4	3		1
w4	1		1	



传统词向量方法

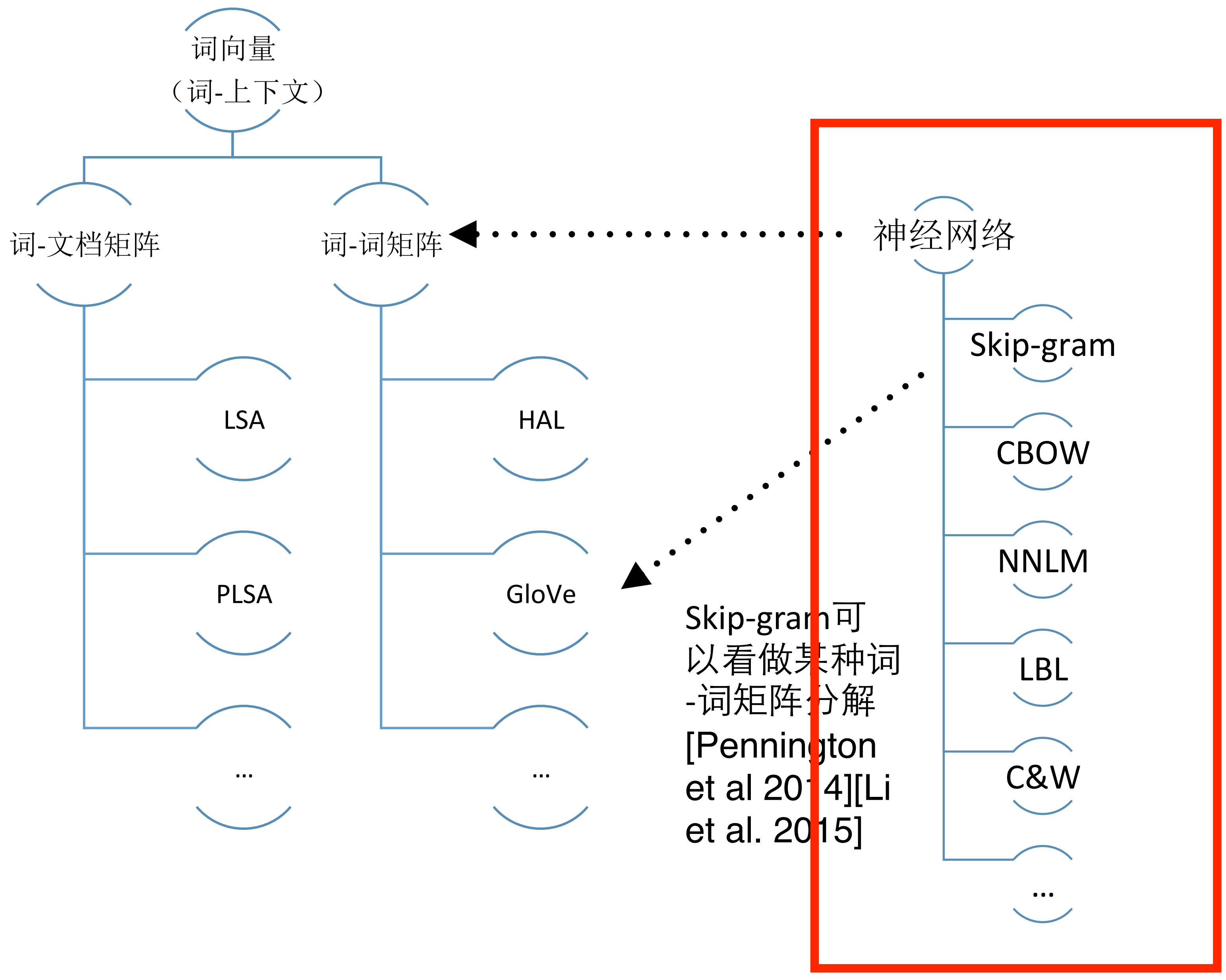
- “词-词”共现矩阵
  - Paradigmatic Relation（聚合/替换关系/二阶关系）: Two words are similar if they tend to appear in similar contexts
  - Use **surrounding words** for building the word space as a paradigmatic use of context [Sahlgren 2006]

I like nature language processing
You like machine learning
We like deep learning

deep→machine

	w0	w1	w2	w3	w4	w5	w6	w7	w8	w9
(w0) I		*1*								
(w1) like	*1*		*1*			*1*	*1*		*1*	*1*
(w2) nature		*1*		*1*						
(w3) language			*1*		*1*					
(w4) processing				*1*						
(w5) You		*1*								
(w6) machine		*1*						*1*		
(w7) learning							*1*			*1*
(w8) We		*1*								
(w9) deep		*1*						*1*		

Map


词向量
（词-上下文）
词-文档矩阵
词-词矩阵
神经网络
LSA
PLSA
...
HAL
GloVe
...
Skip-gram
CBOW
NNLM
LBL
C&W
...
Skip-gram可以看做某种词-词矩阵分解[Pennington et al 2014][Li et al. 2015]

This Talk

- 如何训练得到一组词向量?
- 如何训练得到一组好的词向量?

This Talk

- 如何训练一个好的词向量模型
  - NNLM、LBL、C&W、CBOW、Skip-gram.....
  - 上下文与目标词的关系?
  - 如何表示上下文?

- 如何训练一个好的词向量模型
  - 7个不同任务（相似度、文本分类、NER......）
  - 模型选择（如何对上下文建模）
  - 语料的选择（领域、大小）
  - 参数的选择（迭代次数、词向量的维度）

如何训练得到一组词向量

从语言模型开始

- 目标：计算一个词串的概率

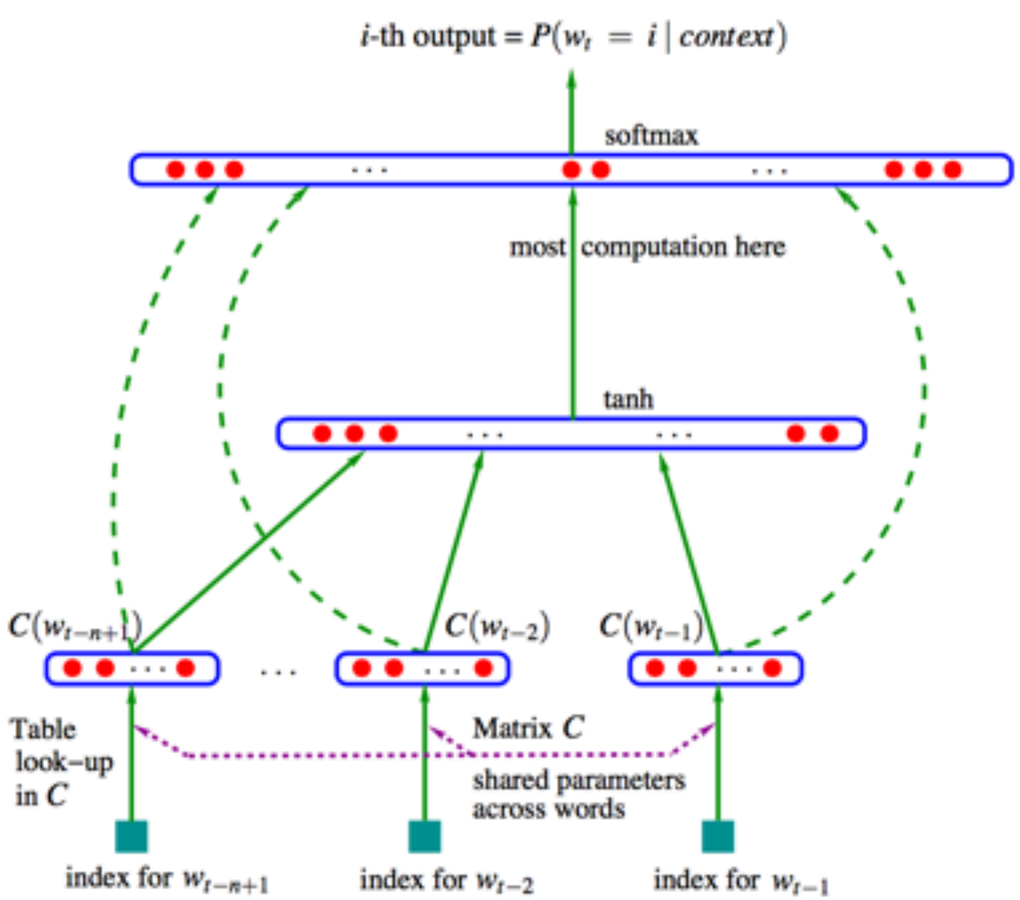
$$
\begin{aligned}
P(S) &= P(w_1, w_2, w_3, \cdots, w_n) \\
&= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)\cdots P(w_n \mid w_1, w_2, w_3, \cdots, w_{n-1}) \\
&= \prod_i P(w_i \mid w_1, w_2, w_3, \cdots, w_{i-1})
\end{aligned}
$$

$$P(w_i \mid w_1, w_2, w_3, \cdots, w_{i-1})$$

$$P(w_i \mid w_1, w_2, w_3, \cdots, w_{i-1}) = \frac{Count(w_1, w_2, w_3, \cdots, w_{i-1}, w_i)}{Count(w_1, w_2, w_3, \cdots, w_{i-1})}$$

NNLM

- Neural Network Language Model [Y.Bengio et al. 2003]



$$L = \frac{1}{T}\sum_t \log f(w_t, w_{t-1}, \cdots, w_{t-n+1}; \theta) + R(\theta).$$

$$f(w_t, w_{t-1}, \cdots, w_{t-n+1}) = \hat{P}(w_t \mid w_{t-1}, \cdots, w_{t-n+1})$$

$$\hat{P}(w_t|w_{t-1}, \cdots w_{t-n+1}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}$$

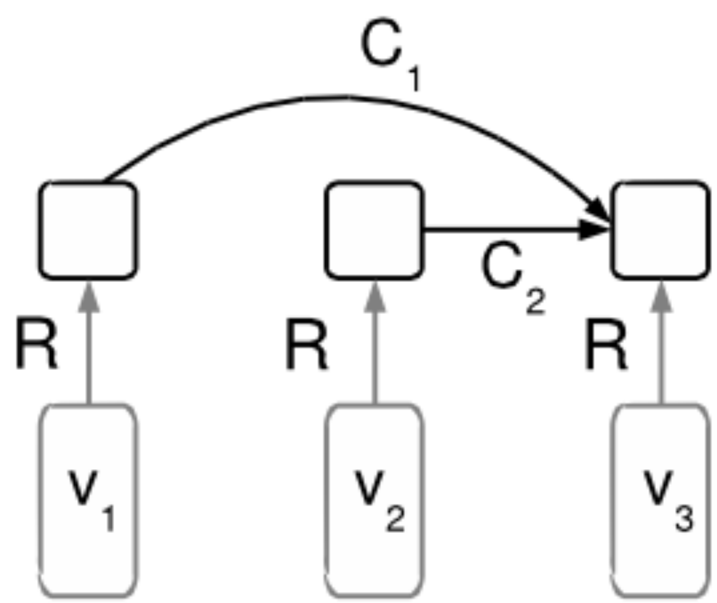
$$y = b + Wx + U\tanh(d + Hx)$$

$$x = (C(w_{t-1}), C(w_{t-2}), \cdots, C(w_{t-n+1}))$$

$$\theta \leftarrow \theta + \varepsilon \frac{\partial \log \hat{P}(w_t|w_{t-1}, \cdots w_{t-n+1})}{\partial \theta}$$

LBL

- Log-bilinear Language Model[A. Mnih & G. Hinton, 2007]



$$P(w_n|w_{1:n-1}) = \frac{1}{Z_c}\exp(-E(w_n;w_{1:n-1}))$$

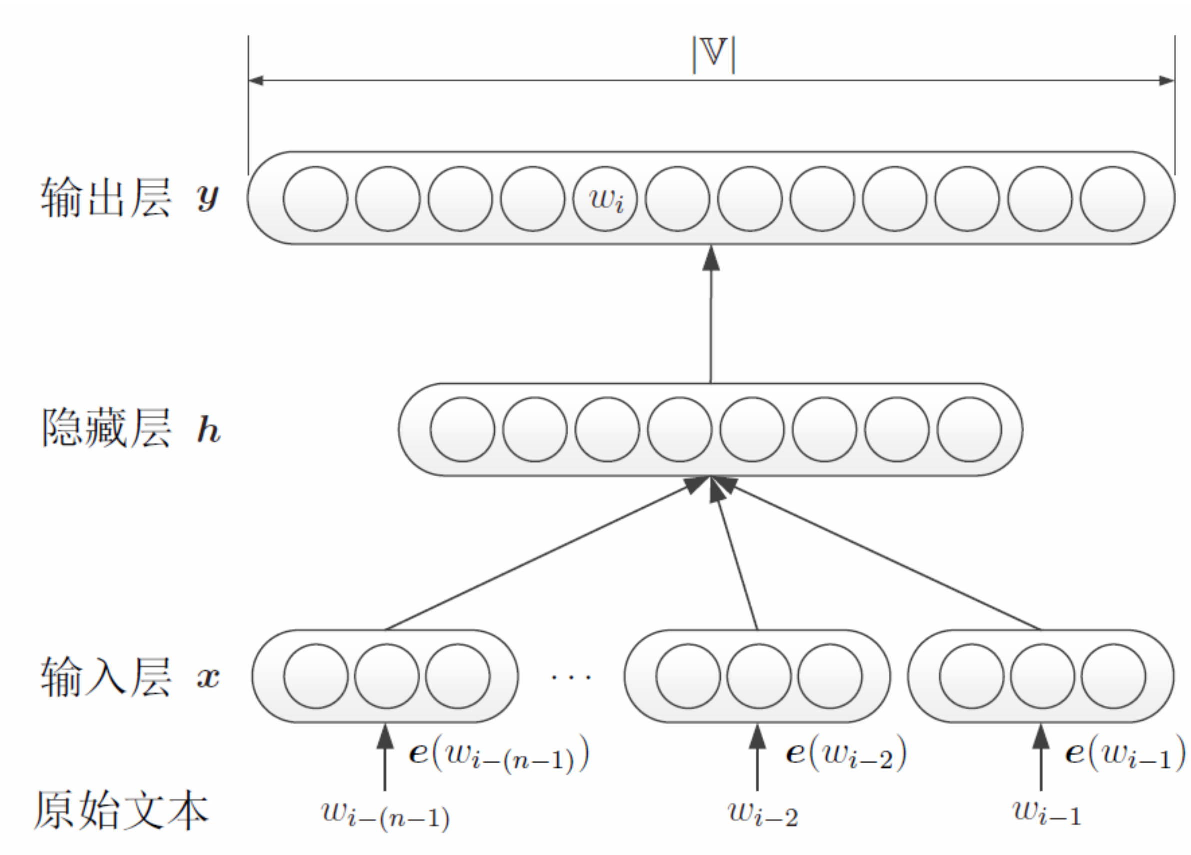
词向量矩阵 词汇表

$$E(w_n;w_{1:n-1}) = -\left(\sum_{i=1}^{n-1} v_i^T R C_i\right) R^T v_n \\ - b_r^T R^T v_n - b_v^T v_n.$$

$$Z_c = \sum_{w_n}\exp(-E(w_n;w_{1:n-1}))$$

LBL vs. NNLM

目标函数： $P(w_n|w_{1:n-1}) = \frac{1}{Z_c}\exp(-E(w_n;w_{1:n-1}))$



NNLM： $y = b + Wx + U\tanh(d + Hx)$

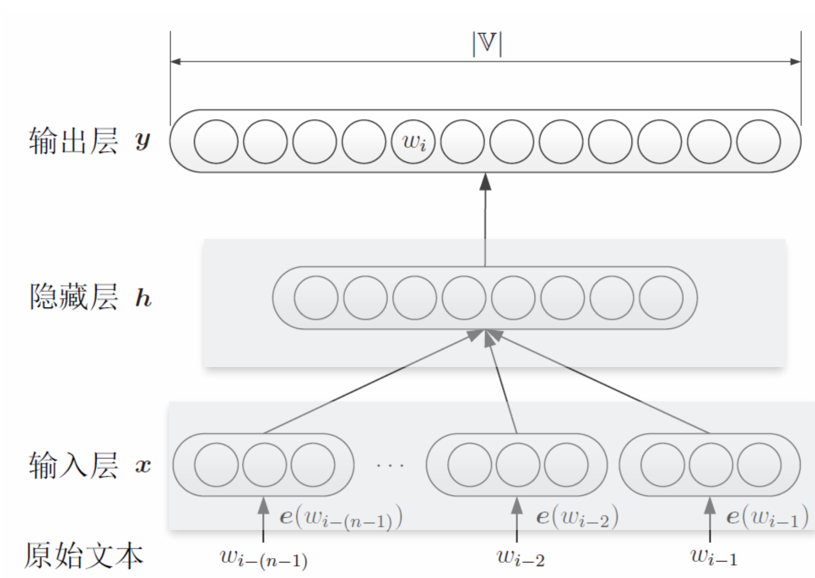
LBL： $$E(w_n;w_{1:n-1}) = -\left(\sum_{i=1}^{n-1} v_i^T RC_i\right) R^T v_n - b_r^T R^T v_n - b_v^T v_n.$$

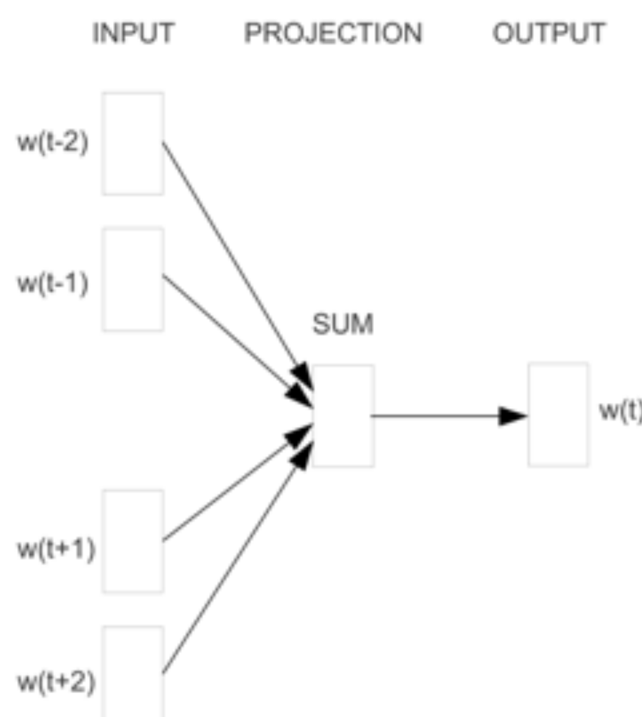
CBOW / Skip-gram

[T. Mikolov et al, ICLR 2013]

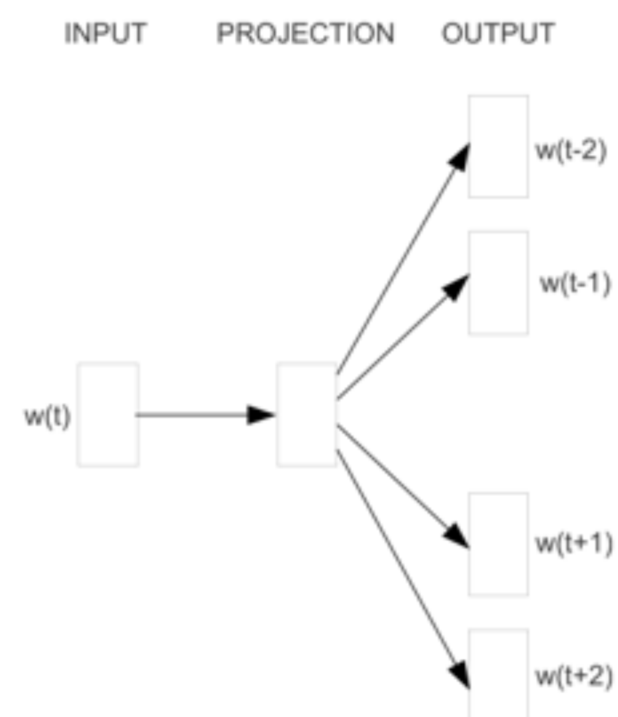
- Word2Vector
  - 去除隐藏层
  - 去除词序

研表究明，汉字序顺并不定一影
阅响读！事证实明了也许当你看
这完句话之后才发字现都乱是的。





Continuous Bag-of-Words



Skip-gram

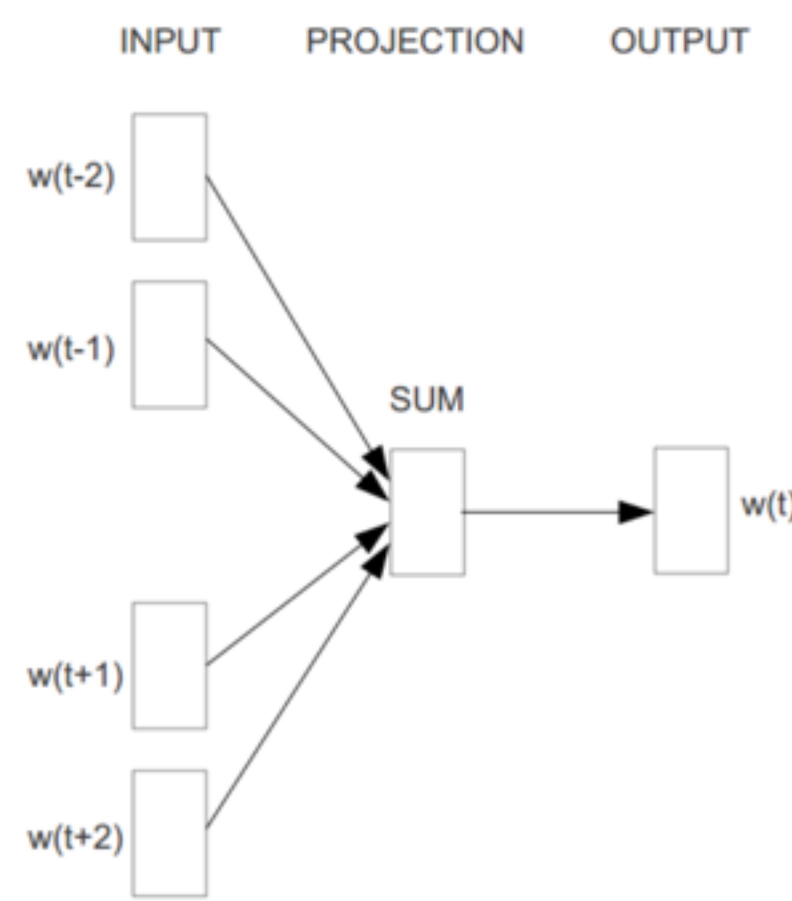
CBOW

- Continued Bag of Words Model

$$\frac{1}{N}\sum_{i=1}^{N} P(w_i \mid w_{i-k}, w_{i-k+1}, \cdots, w_{i-1}, w_{i+1}, \cdots, w_{i+k-1}, w_{i+k})$$

$$P(w_i \mid C_i) = \frac{\exp(v_i'^T v_{C_i})}{\sum_{w_i} \exp(v_i'^T v_{C_i})}$$

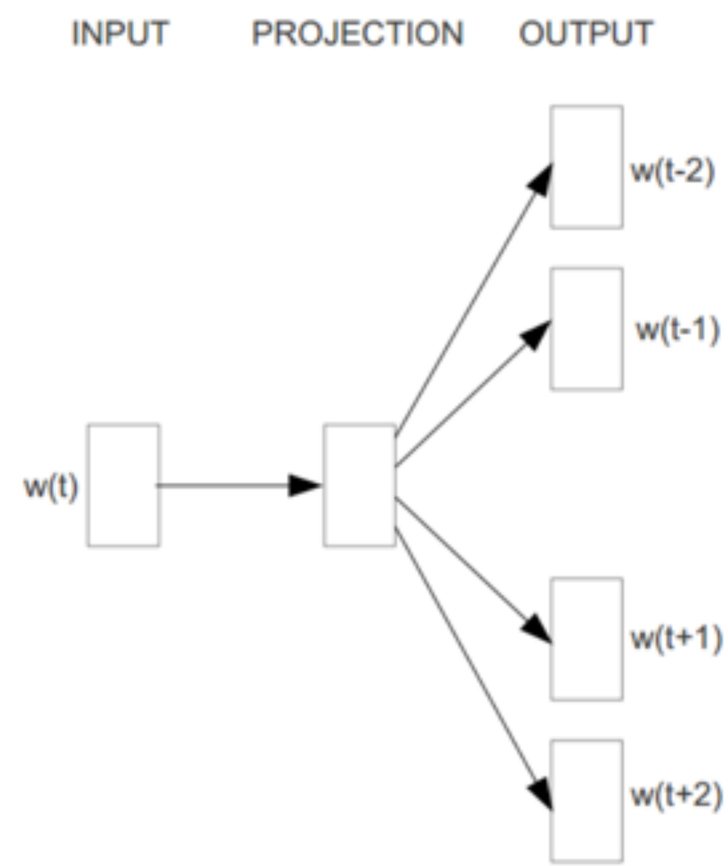
$$v_{C_i} = \sum_{j \in C_i} v_j$$



Skip-Gram

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{-c\le j\le c, j\ne 0} P(w_{i+j} \mid w_i)$$

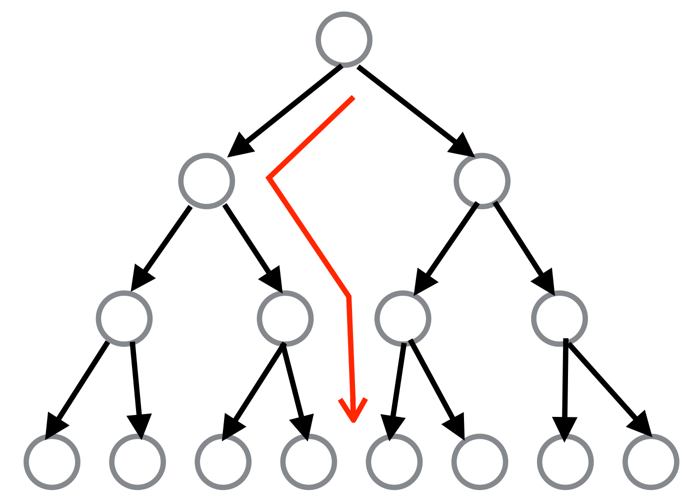
$$P(w_i \mid w_j) = \frac{\exp(v_i'^T v_j)}{\sum_{w_i} \exp(v_i'^T v_j)}$$



加速

- Hierarchical Softmax

$$p(w|w_I) = \prod_{j=1}^{L(w)-1} \sigma\left(\underbrace{[\![n(w,j+1) = \mathrm{ch}(n(w,j))]\!]}_{\text{保证目标词路径的正确}} \cdot {v'_{n(w,j)}}^\top v_{w_I}\right)$$



- Negative Sampling

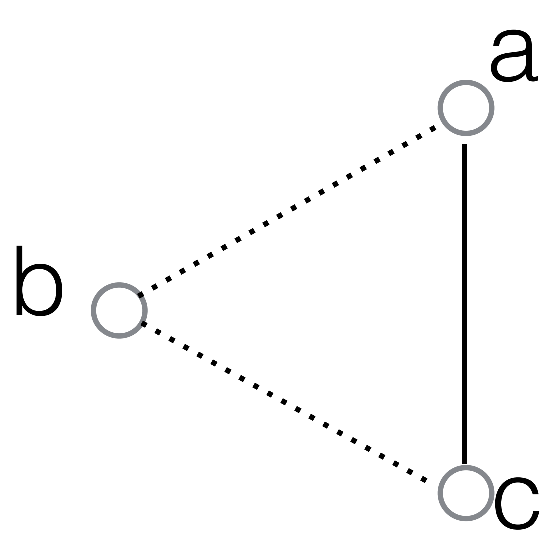
$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\log \sigma({v'_{w_O}}^\top v_{w_I}) + \sum_{i=1}^{k} \underbrace{\mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^\top v_{w_I})\right]}_{\text{按照概率随机抽样}}$$

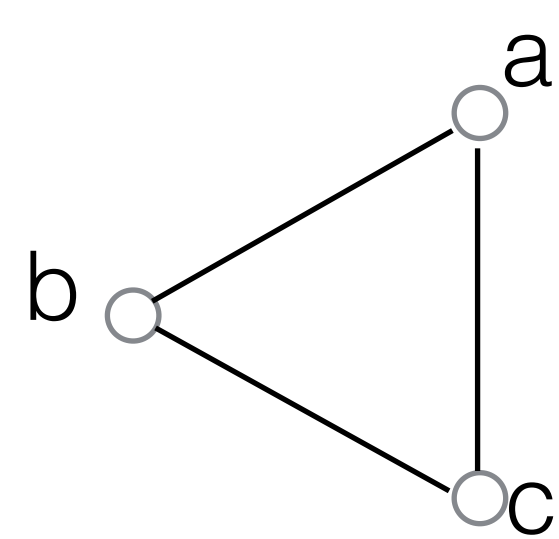
Contextual Vector

$$P(w_i \mid w_j) = \frac{\exp(v_i'^T v_j)}{\sum_{w_i} \exp(v_i'^T v_j)}$$

$$P(w_i \mid w_j) = \frac{\exp(v_i^T v_j)}{\sum_{w_i} \exp(v_i^T v_j)}$$



Paradigmatic Relation



Syntagmatic Relation

Which one should we choose

$\vec{w}_x \quad \vec{c}_x$

- Paradigmatic Relation: $\vec{w}_x$ 或者 $\vec{c}_x$
- Syntagmatic Relation: 两者都要考虑

$\vec{v}_x = \vec{w}_x + \vec{c}_x$

$$cos(x,y) = \frac{\vec{v}_x \cdot \vec{v}_y}{\sqrt{\vec{v}_x \cdot \vec{v}_x}\sqrt{\vec{v}_y \cdot \vec{v}_y}} =$$

$$\frac{(\vec{w}_x + \vec{c}_x) \cdot (\vec{w}_y + \vec{c}_y)}{\sqrt{(\vec{w}_x + \vec{c}_x) \cdot (\vec{w}_x + \vec{c}_x)}\sqrt{(\vec{w}_y + \vec{c}_y) \cdot (\vec{w}_y + \vec{c}_y)}}$$

$$= \frac{\vec{w}_x \cdot \vec{w}_y + \vec{c}_x \cdot \vec{c}_y + \vec{w}_x \cdot \vec{c}_y + \vec{c}_x \cdot \vec{w}_y}{\sqrt{\vec{w}_x^2 + 2\vec{w}_x \cdot \vec{c}_x + \vec{c}_x^2}\sqrt{\vec{w}_y^2 + 2\vec{w}_y \cdot \vec{c}_y + \vec{c}_y^2}}$$

$$= \frac{\vec{w}_x \cdot \vec{w}_y + \vec{c}_x \cdot \vec{c}_y + \vec{w}_x \cdot \vec{c}_y + \vec{c}_x \cdot \vec{w}_y}{2\sqrt{\vec{w}_x \cdot \vec{c}_x + 1}\sqrt{\vec{w}_y \cdot \vec{c}_y + 1}}$$

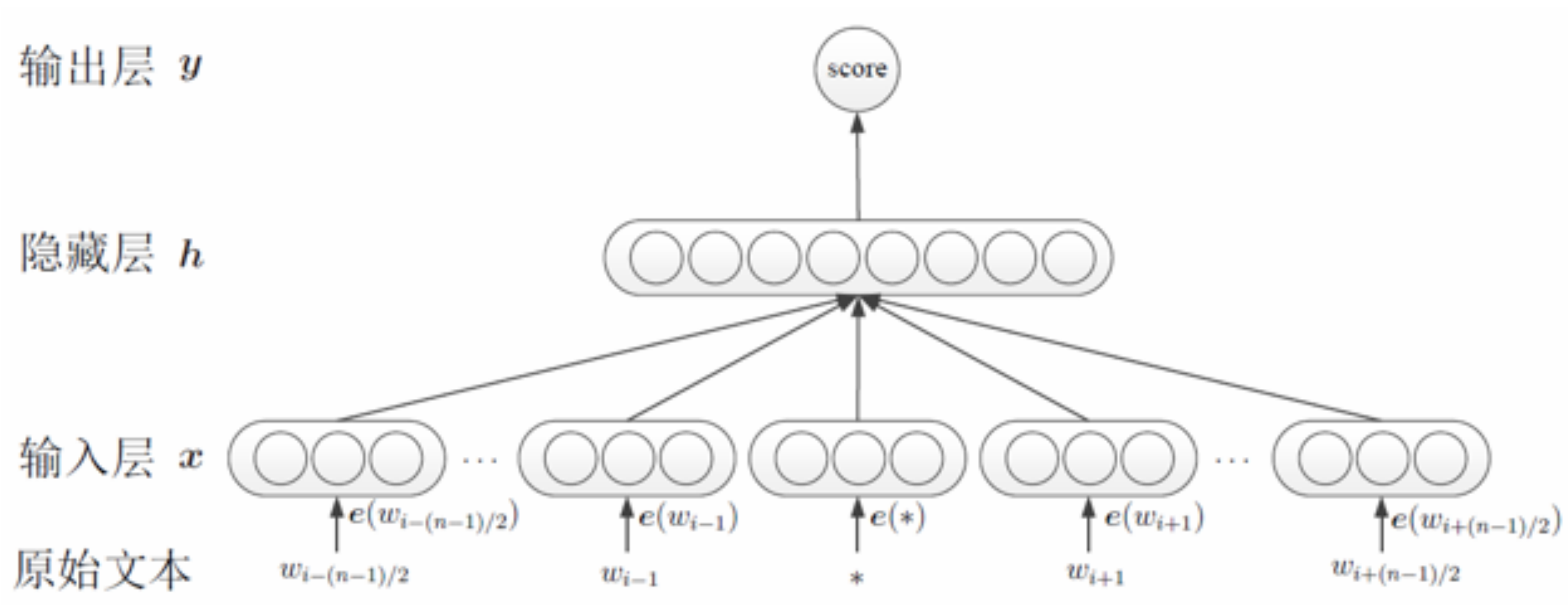
Paradigmatic Relation	$\vec{w}_x \cdot \vec{w}_y$	$\vec{c}_x \cdot \vec{c}_y$
Syntagmatic Relation	$\vec{w}_x \cdot \vec{c}_y$	$\vec{c}_x \cdot \vec{w}_y$

$$sim(x,y) = \frac{sim_2(x,y) + sim_1(x,y)}{\sqrt{sim_1(x,x) + 1}\sqrt{sim_1(y,y) + 1}}$$

[Omer Levy et al, TACL 2015]

C&W [R. Collobert & J. Weston, 2008]

- 目标：词向量



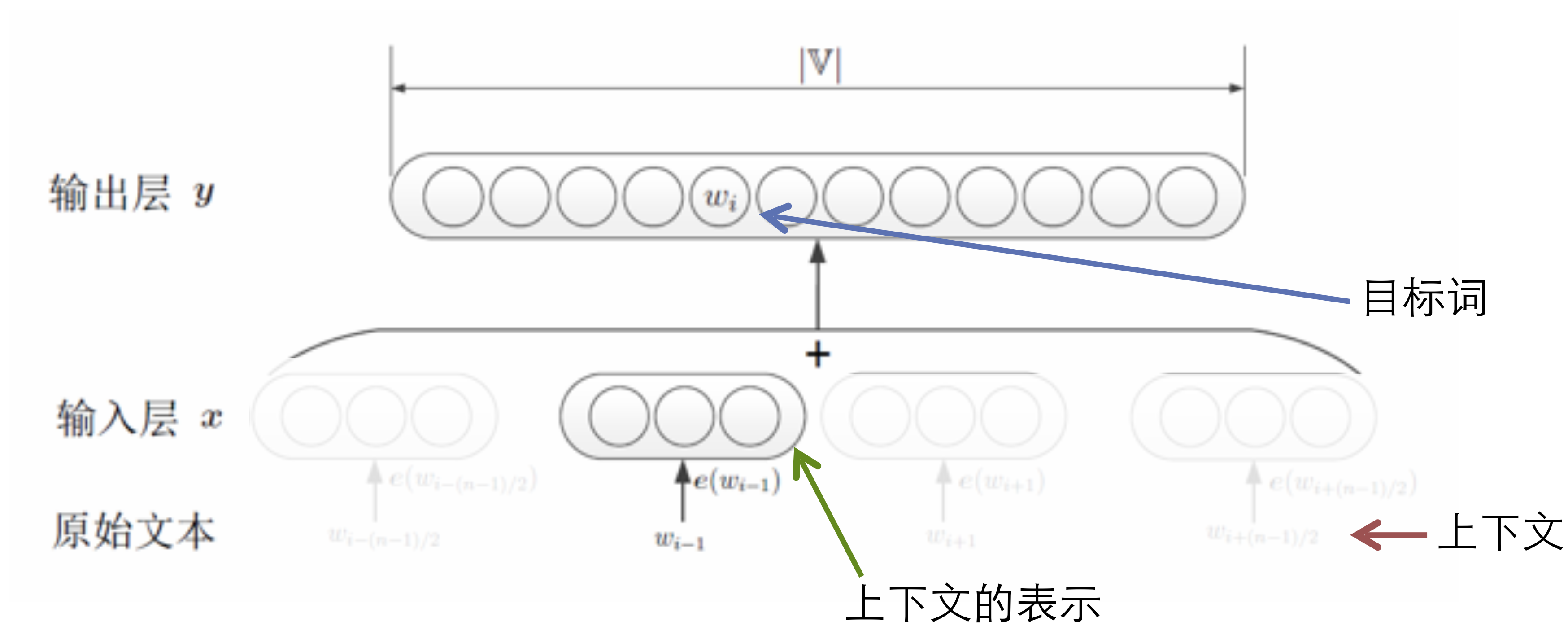
目标函数 $\max(0, 1 - s(w, c) + s(w', c))$

如何训练得到一组好的词向量

模型分析

- 词向量与上下文密切相关
- 两个重要问题
  - 上下文如何表示
  - 上下文与目标词的关系

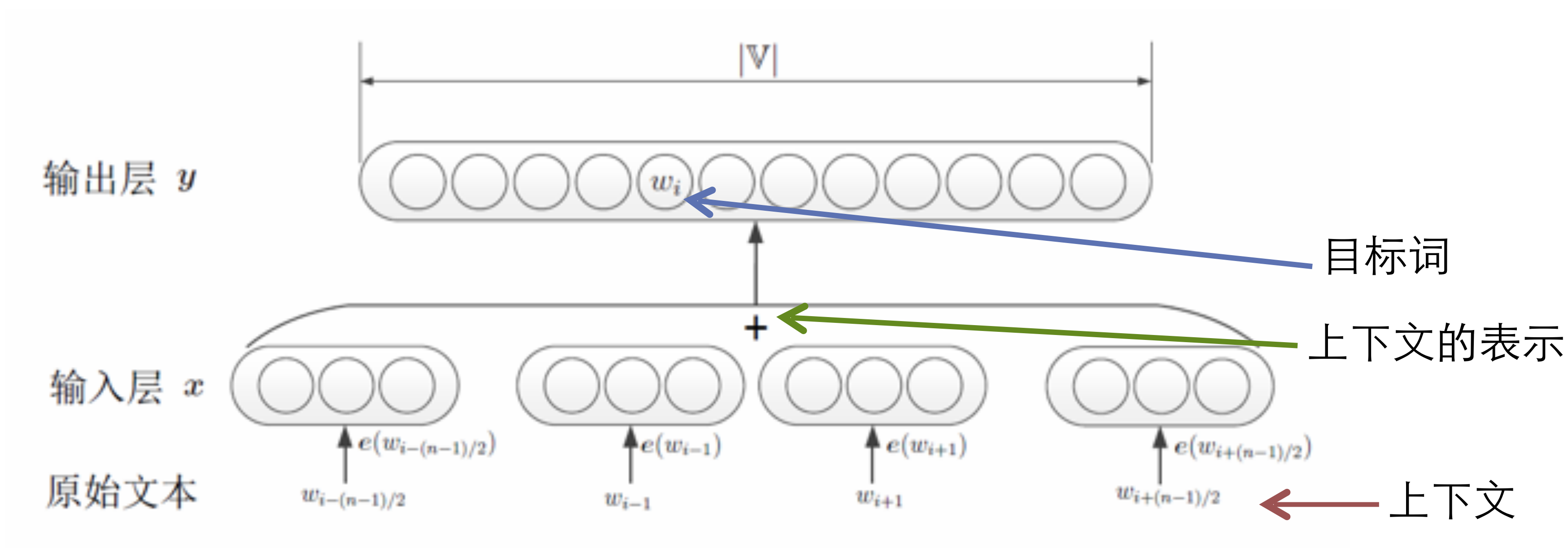
Skip-gram



目标词和上下文的关系: $P(w_i \mid C_i) = P(w_j \mid w_{j+i})$

上下文表示: $e(w_{j+i}), -k \le j \le k, j \ne 0$

CBOW



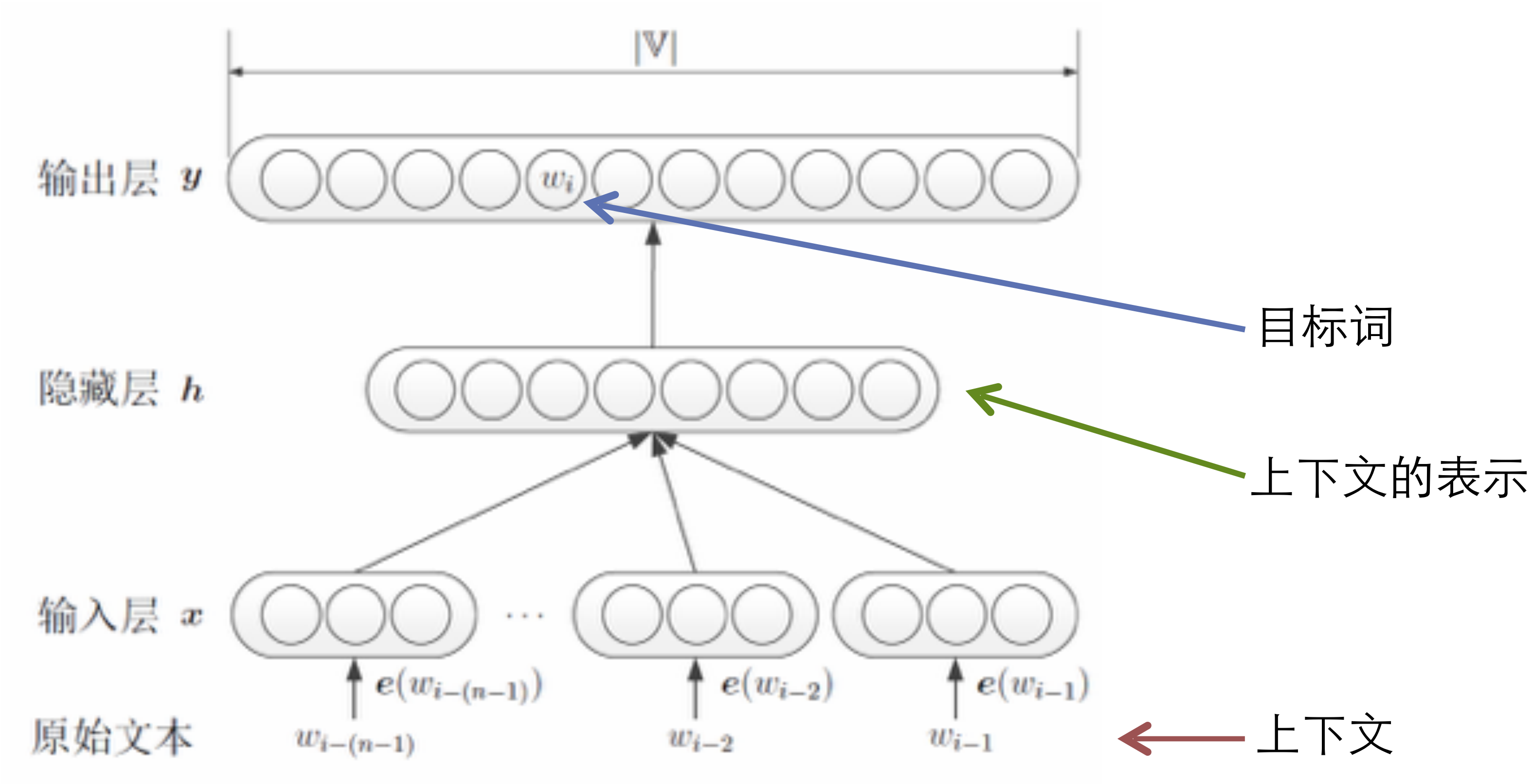
Continuous Bag-of-Words

目标词和上下文的关系： $P(w_i \mid C_i)$

$$= P(w_i \mid w_{i-k}, w_{i-k+1}, \cdots, w_{i-1}, w_{i+1}, \cdots, w_{i+k-1}, w_{i+k})$$

上下文表示： $$\frac{1}{k-1}\left(e(w_{i-\frac{k-1}{2}}) + \cdots + e(w_{i-1}) + e(w_{i+1}) + \cdots + e(w_{i+\frac{k-1}{2}})\right)$$

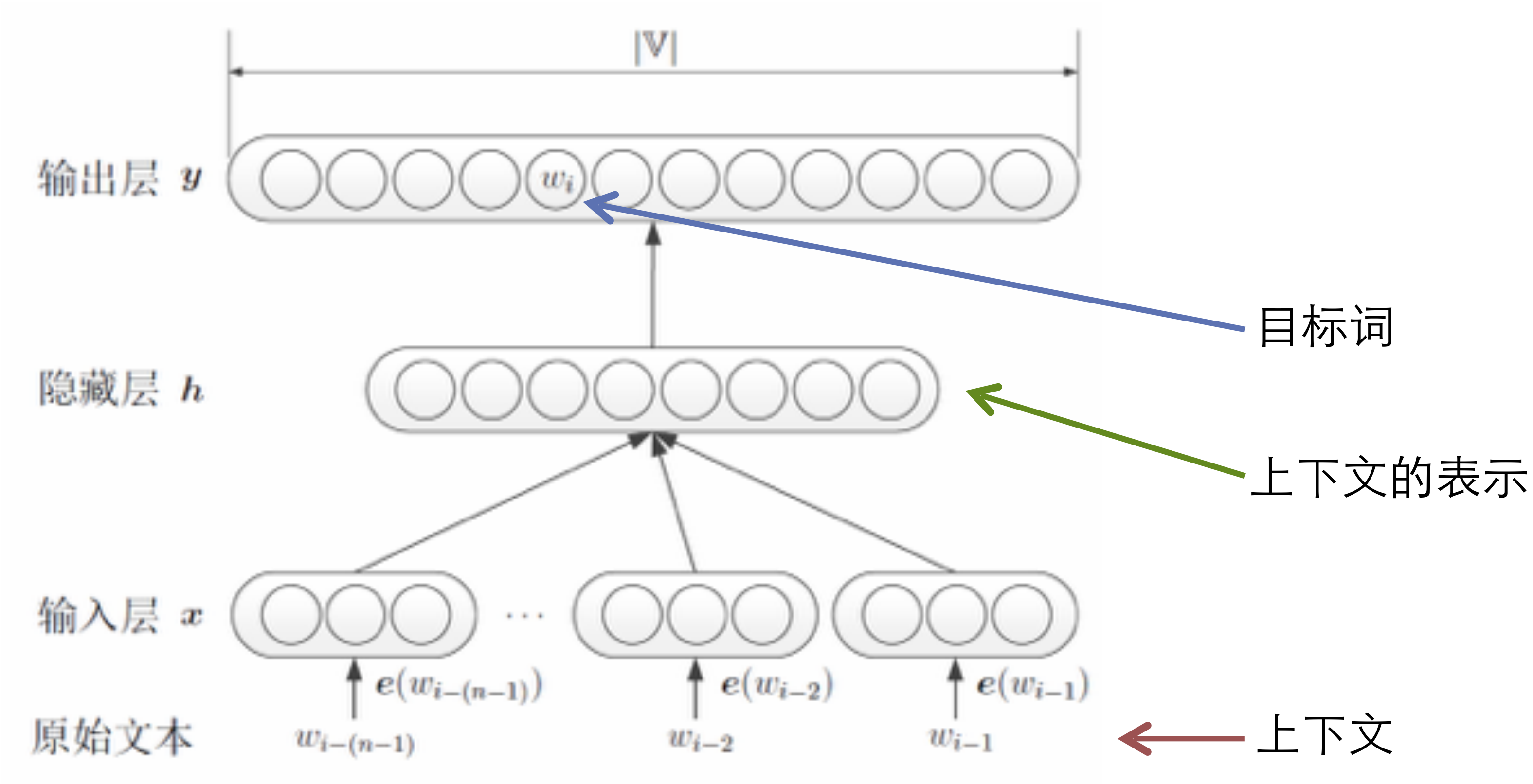
LBL



目标词和上下文的关系：$P(w_i \mid C_i) = P(w_i \mid w_{i-1}, w_{i-2}, \cdots, w_{i-k})$

上下文表示： $H[e(w_1), \cdots, e(w_{n-2}), e(w_{n-1})]$

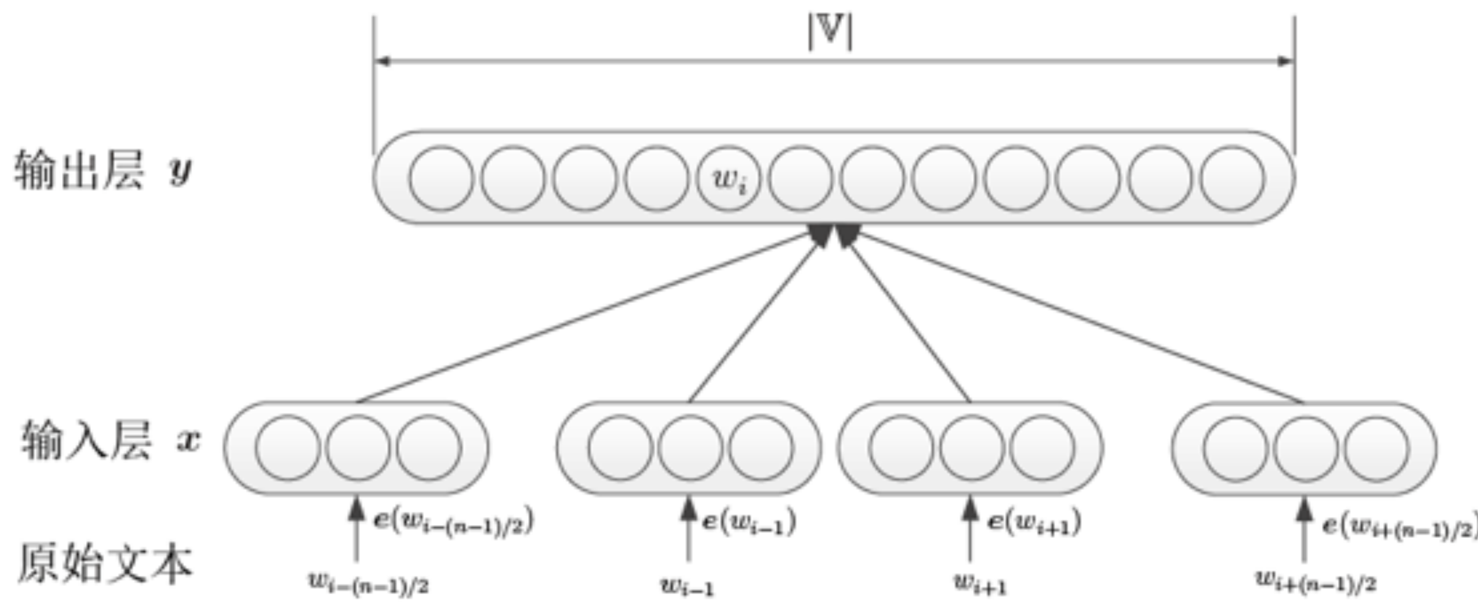
NNLM



目标词和上下文的关系：$P(w_i \mid C_i) = P(w_i \mid w_{i-1}, w_{i-2}, \cdots, w_{i-k})$

上下文表示： $\tanh(d + H[e(w_1), \cdots, e(w_{n-2}), e(w_{n-1})])$

Order(Virtual Model)

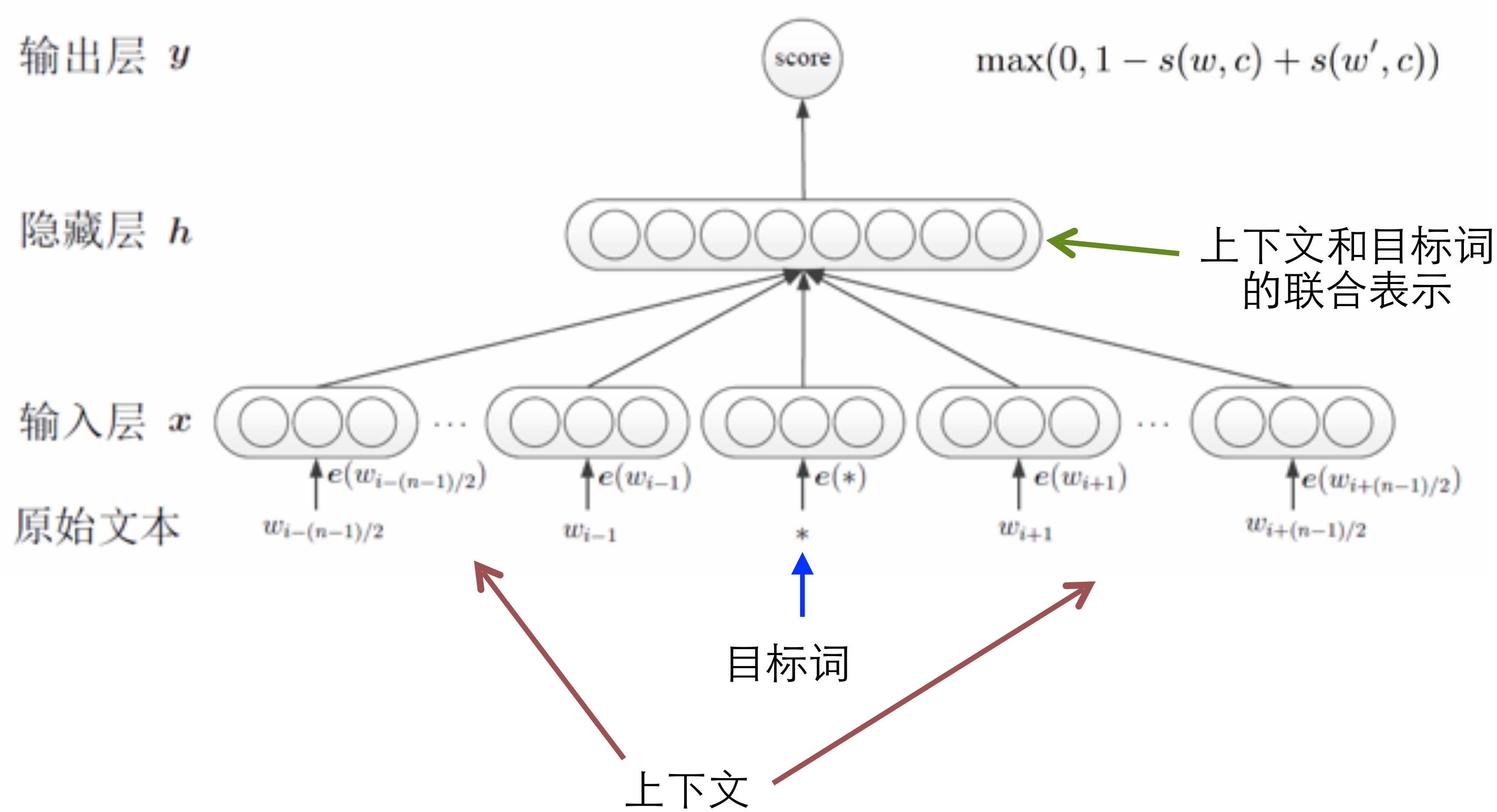


目标词和上下文的关系：$P(w_i | C_i)$

$$= P(w_i | w_{i-k}, w_{i-k+1}, \cdots, w_{i-1}, w_{i+1}, \cdots, w_{i+k-1}, w_{i+k})$$

上下文表示：$[e(w_1), \cdots, e(w_{n-2}), e(w_{n-1})]$

C&W



目标词和上下文的关系： $Score(w_i, C_i)$

上下文表示： $H[e(w_{i-\frac{k-1}{2}}),\cdots,e(w_{i-1}),e(w_i),e(w_{i+1}),\cdots,e(w_{i+\frac{k-1}{2}}))$

模型总结

Model	Relation of w, c	Representation of c
Skip-gram [18]	c predicts w	one of c
CBOW [18]	c predicts w	average
Order	c predicts w	concatenation
LBL [22]	c predicts w	compositionality
NNLM [2]	c predicts w	compositionality
C&W [3]	scores w, c	compositionality

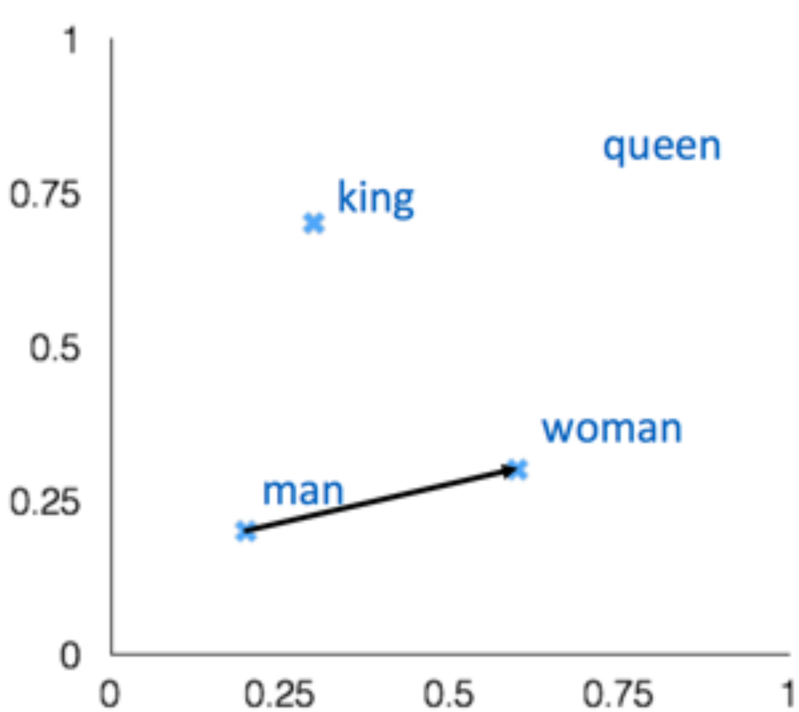
简单

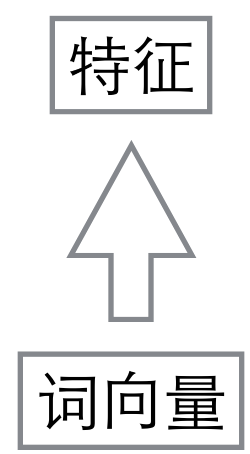
复杂

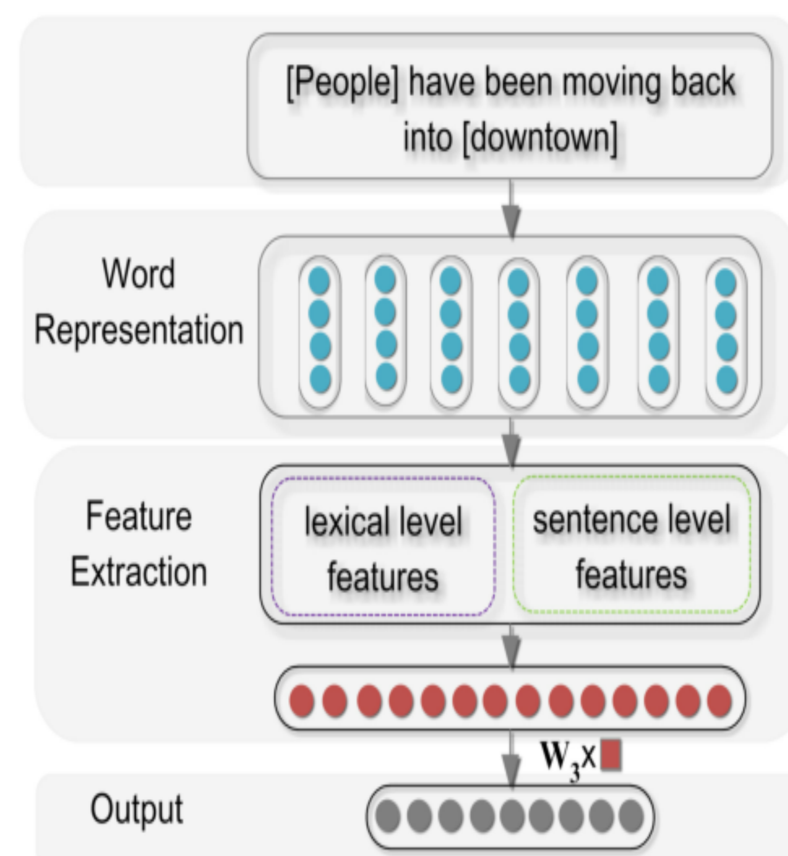
怎样才算是好的词向量

词向量应用

- 语言学应用
- 作为某一任务的特征
- 作为某一任务神经网络模型的初始值







评价任务选择

- 语言学应用
  - 类比任务（syn、sem）
  - 相似度/相关度计算（ws）
  - 同义词（tfl）
- 作为某一任务的特征
  - 情感分类（avg）
  - 命名实体识别（NER）
- 作为某一任务神经网络模型的初始值
  - 情感分类（cnn）
  - 词性标注（pos）

评价任务：类比任务

[Mikolov et al. 2013]

- 语法相似度（syn）10.5k
  - <u>predict</u> – predicting ≈ dance – dancing
- 类比关系（语义）（sem）9k
  - <u>king</u> – queen ≈ man – woman
- 评测
  - man – woman + queen → king
  - predict-dance+dancing →predicting
- 评价指标
  - Accuracy

Model	syn	sem
Random	0.00	0.00
Skip-gram	51.78	**44.80**
CBOW	**55.83**	44.43
Order	55.57	36.38
LBL	45.74	29.12
NNLM	41.41	23.51
C&W	3.13	2.20

评价任务：相似度/相关度

[L. Finkelstein et al., 2013]

- 任务：计算给定词语的相关词语（ws）
  - student, professor 6.81
  - professor, cucumber 0.31
- 数据：WordSim353
- 指标：皮尔逊距离

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Model	ws
Random	0.00
Skip-gram	**63.89**
CBOW	62.21
Order	62.44
LBL	57.86
NNLM	59.25
C&W	46.17

评价任务：同义词

[T. Landauer & S. Dumais, 2013]

- 任务：找给定词语的同义词（tfl）80个选择题

  levied
  A) **imposed** B) believed
  C) requested D) correlated

- 数据：托福考试同义词题
- 指标：Accuracy

Model	tfl
Random	25.00
Skip-gram	76.25
CBOW	**77.50**
Order	**77.50**
LBL	75.00
NNLM	71.25
C&W	47.50

评价任务：文本分类

- 任务：情感分类（avg）
  - 10万条（5万有标注）
  - 25,000 Train, 25,000 Test
- 特征：文档中各词词向量平均值
- 分类模型：Logistic Regression
- 数据：IMDB
- 指标：Accuracy

Model	avg
Random	64.38
Skip-gram	**74.94**
CBOW	74.68
Order	**74.93**
LBL	74.32
NNLM	73.70
C&W	73.26

评价任务：命名实体识别

[Turian et al., 2010]

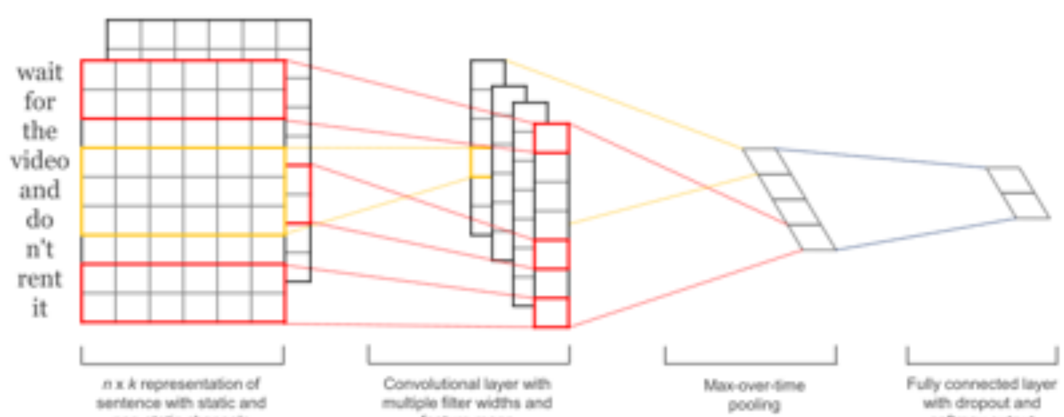
- 任务：NER
- 特征：传统特征[Ratinov 2009]+训练得到的词向量
- 模型：CRFs
- 数据：CoNLL03 shared task
- 指标：F1

Model	ner
Random	84.39
Skip-gram	**88.90**
CBOW	88.47
Order	88.41
LBL	88.69
NNLM	88.36
C&W	88.15

评价任务：情感分类

[Y. Kim, 2014]

- 任务：情感分类，5分类（cnn）
- 模型：Convolutional Neural Network
- 数据：Standford Sentiment Tree Bank
  - 6920 Train, 872 Dev, 1821 Test
- 指标：Accuracy



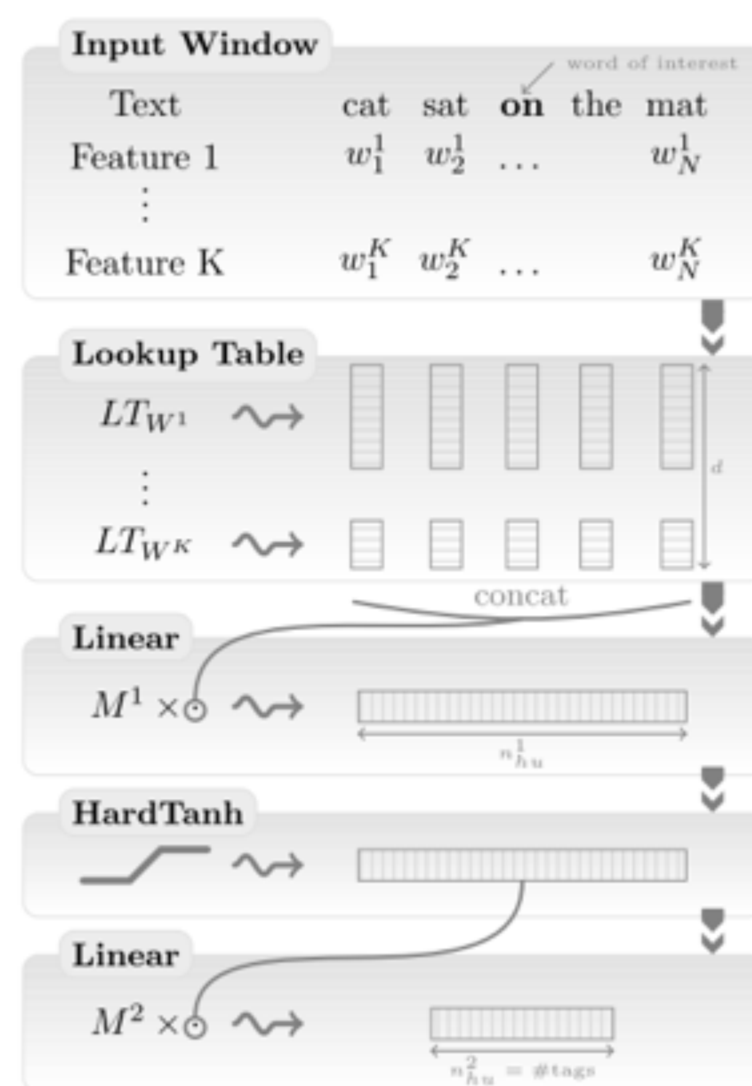
Model	cnn
Random	36.60
Skip-gram	43.84
CBOW	43.75
Order	**44.77**
LBL	43.98
NNLM	44.40
C&W	41.86

评价任务：词性标注

[R. Collobert et al., 2011]

- 任务：标注给定句子中词的词性（pos）数据规模
- 模型：SENNA
- 数据：Wall Street Journal
  - 18,540 Train, 2,824 Dev, 3,229 Test
- 指标：Accuracy

Model	pos
Random	95.41
Skip-gram	96.57
CBOW	96.63
Order	**96.76**
LBL	**96.77**
NNLM	96.73
C&W	96.66



实验设置

- Corpus
  - Wiki:100M, 1.6B
  - NYT: 100M, 1.2B
  - W&N: 10M, 100M, 1B, 2.8B
  - IMDB: 13M
- Parameters
  - Dimension: 10, 20, 50, 100, 200
  - Window size: 5

结果

Model	syn	sem	ws	tfl	avg	ner	cnn	pos
Random	0.00	0.00	0.00	25.00	64.38	84.39	36.60	95.41
Skip-gram	51.78	**44.80**	**63.89**	76.25	**74.94**	**88.90**	43.84	96.57
CBOW	**55.83**	44.43	62.21	**77.50**	74.68	88.47	43.75	96.63
Order	55.57	36.38	62.44	**77.50**	**74.93**	88.41	**44.77**	**96.76**
LBL	45.74	29.12	57.86	75.00	74.32	88.69	43.98	**96.77**
NNLM	41.41	23.51	59.25	71.25	73.70	88.36	44.40	96.73
C&W	3.13	2.20	46.17	47.50	73.26	88.15	41.86	96.66

语言学特性　　作为特征　　作为网络输入

问题：不同任务间很难进行公平比较

评价指标：效果增益率

- Performance Gain Ratio

$$PGR(a,b) = \frac{p_a - p_{rand}}{p_b - p_{rand}}$$

Model	syn	sem	ws	tfl	avg	ner	cnn	pos
Random	0.00	0.00	0.00	25.00	64.38	84.39	36.60	95.41

$$PGR(a,\max) = \frac{p_a - p_{rand}}{p_{\max} - p_{rand}}$$

结果

Model	syn	sem	ws	tfl	avg	ner	cnn	pos
Random	0.00	0.00	0.00	25.00	64.38	84.39	36.60	95.41
Skip-gram	51.78	**44.80**	**63.89**	76.25	**74.94**	**88.90**	43.84	96.57
CBOW	**55.83**	44.43	62.21	**77.50**	74.68	88.47	43.75	96.63
Order	55.57	36.38	62.44	**77.50**	**74.93**	88.41	**44.77**	**96.76**
LBL	45.74	29.12	57.86	75.00	74.32	88.69	43.98	**96.77**
NNLM	41.41	23.51	59.25	71.25	73.70	88.36	44.40	96.73
C&W	3.13	2.20	46.17	47.50	73.26	88.15	41.86	96.66

Model	syn	sem	ws	tfl	avg	ner	cnn	pos
Skip-gram	93	**100**	**100**	98	**100**	**100**	89	85
CBOW	**100**	99	97	**100**	98	90	88	90
Order	**100**	81	98	**100**	**100**	89	**100**	99
LBL	82	65	91	95	94	95	90	**100**
NNLM	74	52	93	88	88	88	95	97
C&W	6	5	72	43	84	83	64	92

上下文和目标词的关系

Model	syn	sem	ws	tfl	avg	ner	cnn	pos
Skip-gram	93	100	100	98	100	100	89	85
CBOW	100	99	97	100	98	90	88	90
Order	100	81	98	100	100	89	100	99
LBL	82	65	91	95	94	95	90	100
NNLM	74	52	93	88	88	88	95	97
C&W	6	5	72	43	84	83	64	92

上下文预测目标词

上下文、目标词 联合打分

C&W: Syntagmatic Relation

Skip-gram, CBOW, Order, LBL, NNLM: Paradigmatic Relation

上下文和目标词的关系

Model	syn	sem	ws	tfl	avg	ner	cnn	pos
Skip-gram	93	100	100	98	100	100	89	85
CBOW	100	99	97	100	98	90	88	90
Order	100	81	98	100	100	89	100	99
LBL	82	65	91	95	94	95	90	100
NNLM	74	52	93	88	88	88	95	97
C&W	6	5	72	43	84	83	64	92

Model	Monday	commonly
CBOW	Thursday	generically
	Friday	colloquially
	Wednesday	popularly
	Tuesday	variously
	Saturday	Commonly
C&W	8:30	often
	12:50	generally
	1PM	previously
	4:15	have
	mid-afternoon	are

paradigmatic relation

syntagmatic relation

上下文表示

Model	syn	sem	ws	tfl	avg	ner	cnn	pos	
Skip-gram	93	100	100	98	100	100	89	85	3+2
CBOW	100	99	97	100	98	90	88	90	
Order	100	81	98	100	100	89	100	99	
LBL	82	65	91	95	94	95	90	100	1+2
NNLM	74	52	93	88	88	88	95	97	
C&W	6	5	72	43	84	83	64	92	

上下文表示

简单

Model	10M	100M	1B	2.8B
Skip-gram	4+2	4+2	2+2	3+2
CBOW	1+1	3+3	4+1	4+1
Order	0+2	1+2	2+3	3+3
LBL	0+2	0+2	0+2	1+2
NNLM	0+2	0+3	0+3	0+2

复杂

W&N

小语料时，简单的上下文表示有效果
随着语料规模的增大，相对复杂的语料展现较好的结果

语料规模的影响

- 同领域语料，越大越好

Corpus	syn	sem	ws	tfl	avg	ner	cnn	pos
NYT 1.2B	93	52	90	98	50	76	85	96
100M	76	30	88	93	46	77	83	86
Wiki 1.6B	92	**100**	**100**	93	51	**100**	86	94
100M	74	65	98	93	47	88	90	83
W&N 2.8B	**100**	89	95	93	50	97	91	**100**
1B	98	87	95	**100**	48	98	90	98
100M	79	63	97	96	51	85	92	86
10M	29	27	76	60	42	49	77	42
IMDB 13M	32	21	55	82	**100**	26	**100**	-13

CBOW

语料规模的影响

- syn任务，语料越大越好

Corpus	syn	sem	ws	tfl	avg	ner	cnn	pos
NYT 1.2B	93	52	90	98	50	76	85	96
100M	76	30	88	93	46	77	83	86
Wiki 1.6B	92	**100**	**100**	93	51	**100**	86	94
100M	74	65	98	93	47	88	90	83
W&N 2.8B	**100**	89	95	93	50	97	91	**100**
1B	98	87	95	**100**	48	98	90	98
100M	79	63	97	96	51	85	92	86
10M	29	27	76	60	42	49	77	42
IMDB 13M	32	21	55	82	**100**	26	**100**	-13

CBOW

语料领域的影响

- 对于语义相似度任务（sem、ws），维基百科具有优势

Corpus	syn	sem	ws	tfl	avg	ner	cnn	pos
NYT 1.2B	93	52	90	98	50	76	85	96
100M	76	30	88	93	46	77	83	86
Wiki 1.6B	92	**100**	**100**	93	51	**100**	86	94
100M	74	65	98	93	47	88	90	83
W&N 2.8B	**100**	89	95	93	50	97	91	**100**
1B	98	87	95	**100**	48	98	90	98
100M	79	63	97	96	51	85	92	86
10M	29	27	76	60	42	49	77	42
IMDB 13M	32	21	55	82	**100**	26	**100**	-13

CBOW

语料领域的影响

- 领域相关任务：利用领域内语料训练效果好

Corpus	movie	Sci-Fi	season
IMDB	film this it thing miniseries	SciFi sci-fi fi Sci SF	episode seasons installment episodes series
W&N	film big-budget movies live-action low-budget	Nickelodeon Cartoon PBS SciFi TV	half-season seasons homestand playoffs game

				tfl	avg	ner	cnn	pos
				98	50	76	85	96
				93	46	77	83	86
				93	51	**100**	86	94
				93	47	88	90	83
				93	50	97	91	**100**
				100	48	98	90	98
100M	79	63	97	96	51	85	92	86
10M	29	27	76	60	42	49	77	42
IMDB 13M	32	21	55	82	**100**	26	**100**	-13

CBOW

语料领域和大小哪一个更重要

- 情感分类

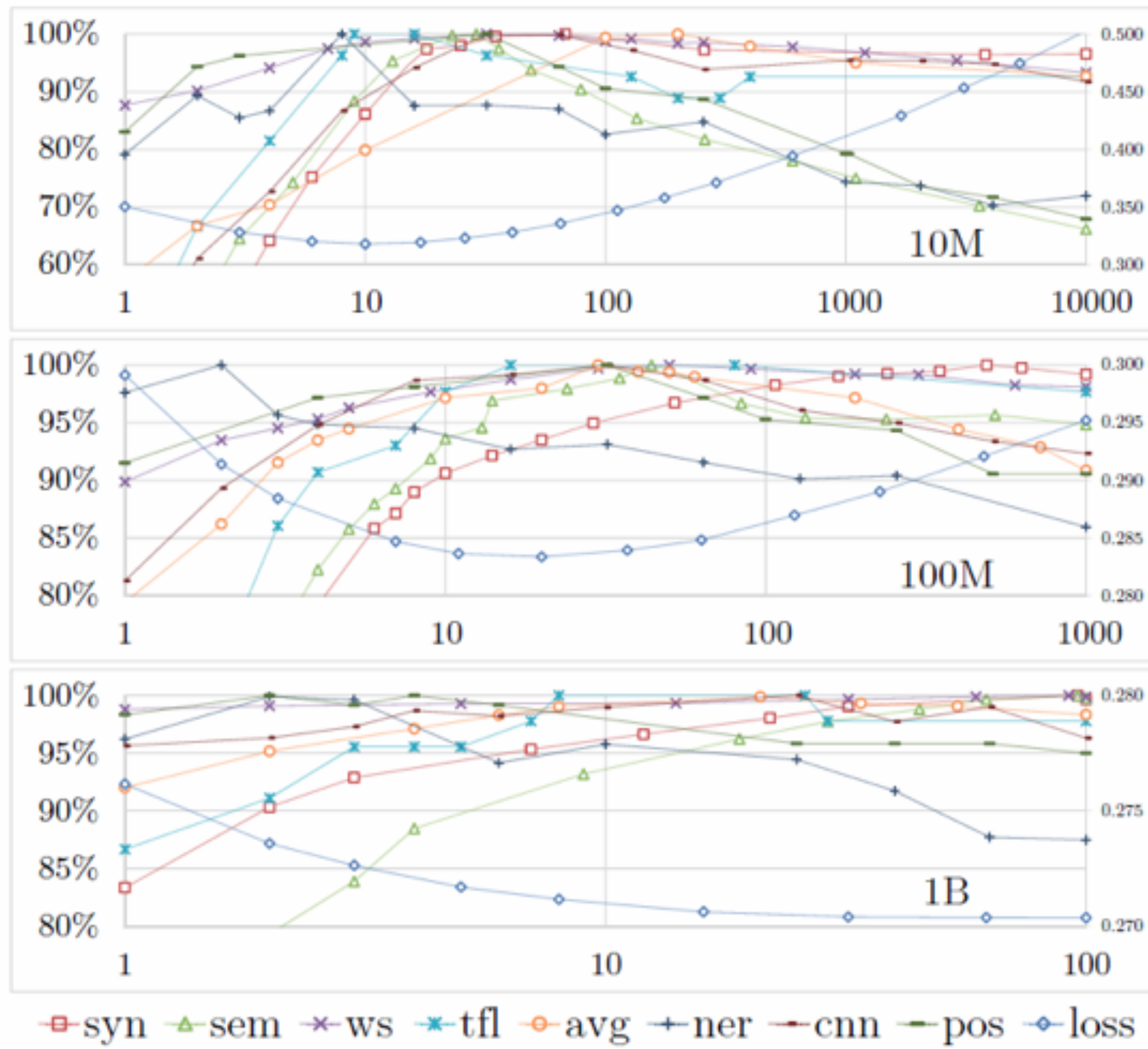
W&N \ IMDB	20%	40%	60%	80%	100%
+0%	91	94	100	100	100
+20%	79	87	91	96	99
+40%	68	86	88	92	98
+60%	65	79	85	88	93
+80%	64	75	84	87	92
+100%	64	70	83	86	88

CBOW

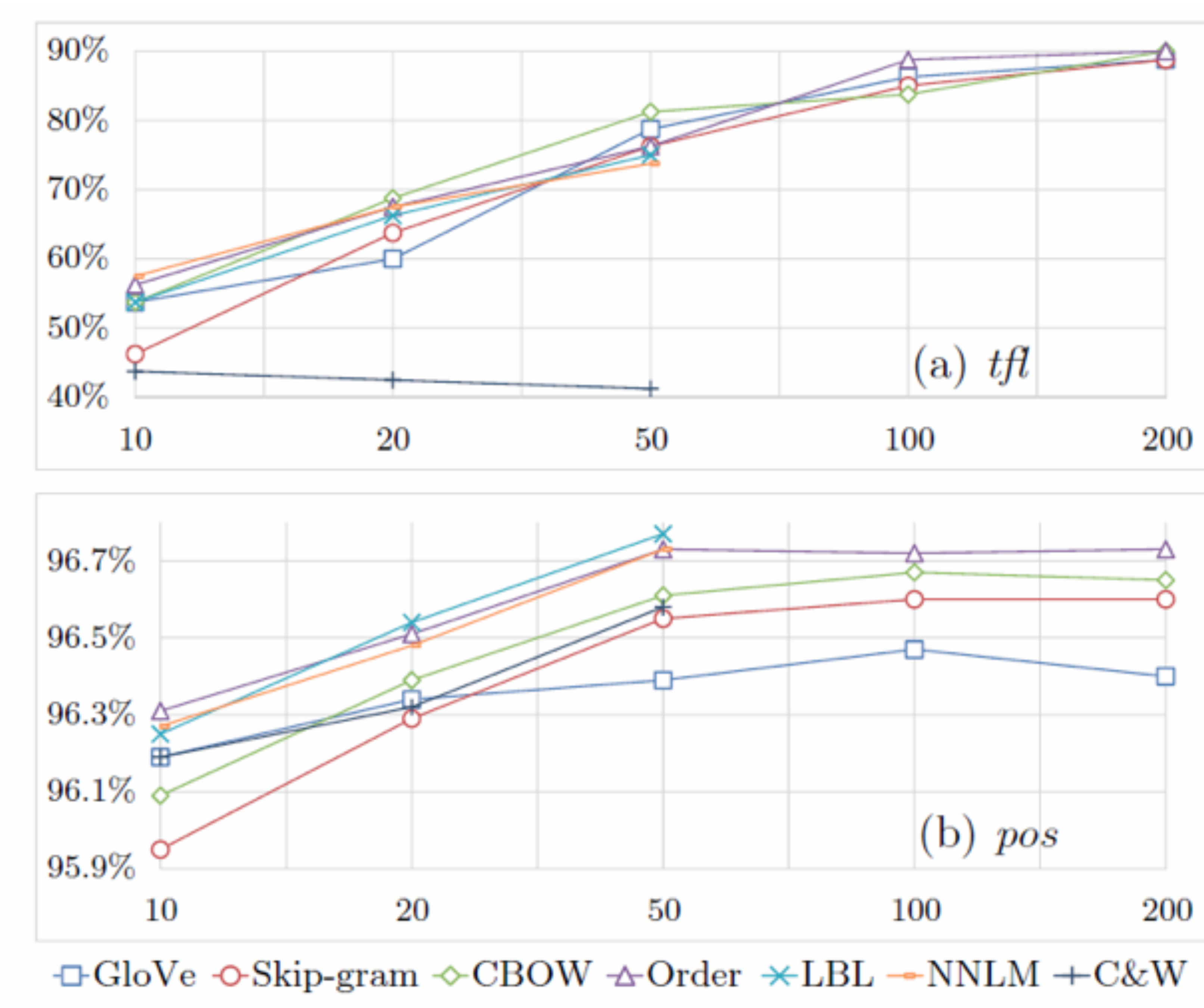
领域更加重要

训练参数：Iteration Number

- Early stop



训练参数：Dimension



总结

- 没有最好，只有适合
  - 适合任务，用（任务相关）领域内语料训练
- 确定合适领域的语料之后，语料越大越好
- 大语料（数据丰富），使用复杂模型（NNLM、C&W）
- 小语料（数据稀疏），使用简单模型（Skip-gram）
- 使用任务的验证集，而非词向量的验证集
- 词向量维度建议50以上
- 注意区分Syntagmatic(组合/一阶)关系和Paradigmatic(替换/二阶)关系

未来

- 跨领域训练词向量
  - ACL2015: Unsupervised Cross-Domain Word Representation Learning

W&N \ IMDB	20%	40%	60%	80%	100%
+0%	91	94	100	100	100
+20%	79	87	91	96	99
+40%	68	86	88	92	98
+60%	65	79	85	88	93
+80%	64	75	84	87	92
+100%	64	70	83	86	88

未来

- 词向量和已有人工标注相结合
  - ACL2015 Best Paper:AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes
    - 联合学习word、synset、lexemes的向量表示
  - EMNLP2014：Knowledge Graph and Text Jointly Embeddings
    - 联合学习词、知识库的向量表示

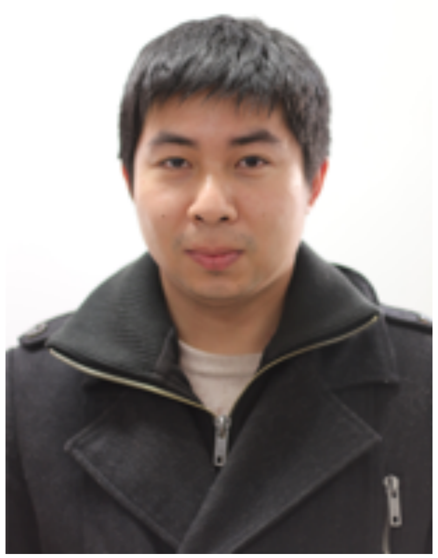
本项工作

- Siwei Lai, Kang Liu, Liheng Xu, Jun Zhao. How to Generate a Good Word Embedding? In http://arxiv.org/abs/1507.05523

- Code: https://github.com/licstar/compare

- 中文导读：http://licstar.net/archives/620

来斯惟



徐立恒

赵军

Reference

- [1] M. Baroni, G. Dinu, and G. Kruszewski. Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In ACL, 2014.
- [2] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A Neural Probabilistic Language Model. JMLR, 2003.
- [3] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In ICML, 2008.
- [4] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. JMLR, 12, 2011.
- [5] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. JMLR, 2011.
- [6] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. Why does unsupervised pre-training help deep learning? JMLR, 2010.
- [7] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. TOIS, 2002.
- [8] M. U. Gutmann and A. Hyvarinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. JMLR, 2012.
- [9] Z. S. Harris. Distributional structure. Word, 1954.
- [10] K. M. Hermann. Distributed Representations for Compositional Semantics. PhD thesis, University of Oxford, 2014.
- [11] Y. Kim. Convolutional neural networks for sentence classification. In EMNLP, 2014.

Reference

- [12] T. K. Landauer. On the computational basis of learning and cognition: Arguments from lsa. Psychology of learning and motivation, 2002.
- [13] T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 1997.
- [14] R. Lebret and R. Collobert. Word embeddings through hellinger pca. In EACL, 2014.
- [15] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In NIPS, 2014.
- [16] O. Levy, Y. Goldberg, and I. Dagan. Improving distributional similarity with lessons learned from word embeddings. TACL, 2015.
- [17] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In ACL, 2011.
- [18] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. ICLR Workshop Track, 2013.
- [19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In NIPS, 2013.
- [20] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In NAACL-HLT, 2013.
- [21] D. Milajevs, D. Kartsaklis, M. Sadrzadeh, and M. Purver. Evaluating neural word representations in tensor-based compositional settings. In EMNLP, 2014.
- [22] A. Mnih and G. Hinton. Three new graphical models for statistical language modelling. In ICML, 2007.
- [23] A. Mnih and G. E. Hinton. A scalable hierarchical distributed language model. In NIPS, 2009.

Reference

- [23] A. Mnih and G. E. Hinton. A scalable hierarchical distributed language model. In NIPS, 2009.
- [24] A. Mnih and K. Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In NIPS, 2013.
- [25] F. Morin and Y. Bengio. Hierarchical probabilistic neural network language model. In AISTATS, 2005.
- [26] J. Pennington, R. Socher, and C. D. Manning. GloVe : Global Vectors for Word Representation. In EMNLP, 2014.
- [27] L. Prechelt. Early stopping-but when? Neural Networks: Tricks of the trade, 1998.
- [28] R. Rapp. The computation of word associations: comparing syntagmatic and paradigmatic approaches. In Coling, 2002.
- [29] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In CoNLL, 2009.
- [30] M. Sahlgren. The Word-Space Model. PhD thesis, Gothenburg University, 2006.
- [31] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In EMNLP, 2013.
- [32] P. Stenetorp, H. Soyer, S. Pyysalo, S. Ananiadou, and T. Chikayama. Size (and domain) matters: Evaluating semantic word space representations for biomedical text. In SMBM, 2012.
- [33] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In NAACL-HLT, 2003.
- [34] J. Turian, L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. In ACL, 2010.

谢谢! Q&A!