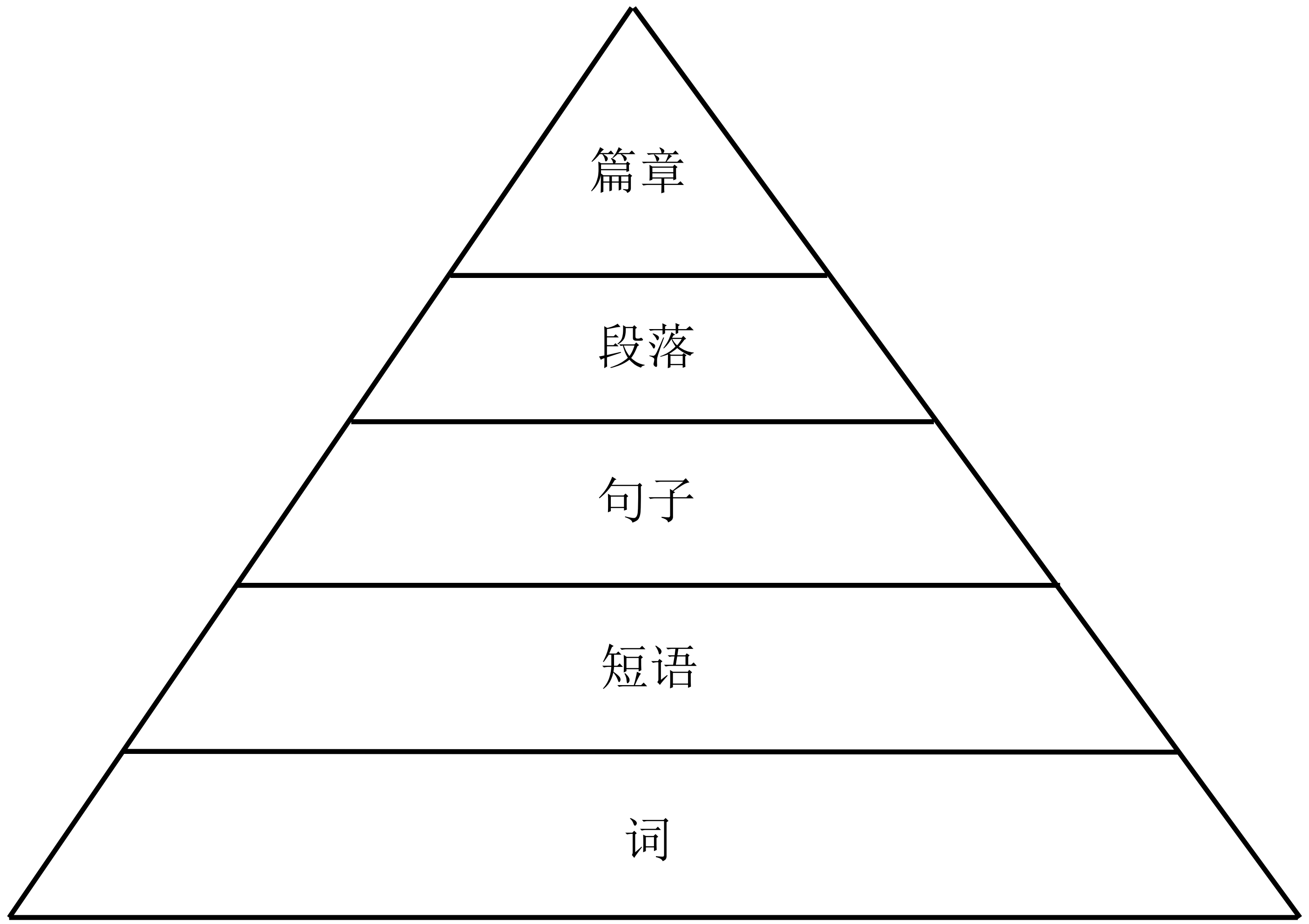
词向量
Word Embedding

词向量 Word Embedding

- 词是最基础的语言单元



词是自然语言处理的基础

- 文本分类

诺基亚5800屏幕很好，操作也很方便，通话质量也不错，

诺基亚	*1*
5800	*1*
屏幕	*1*
很好	*1*
操作	*1*
也	*2*
很	*1*
方便	*1*
通话	*1*
质量	*1*
不错	*1*

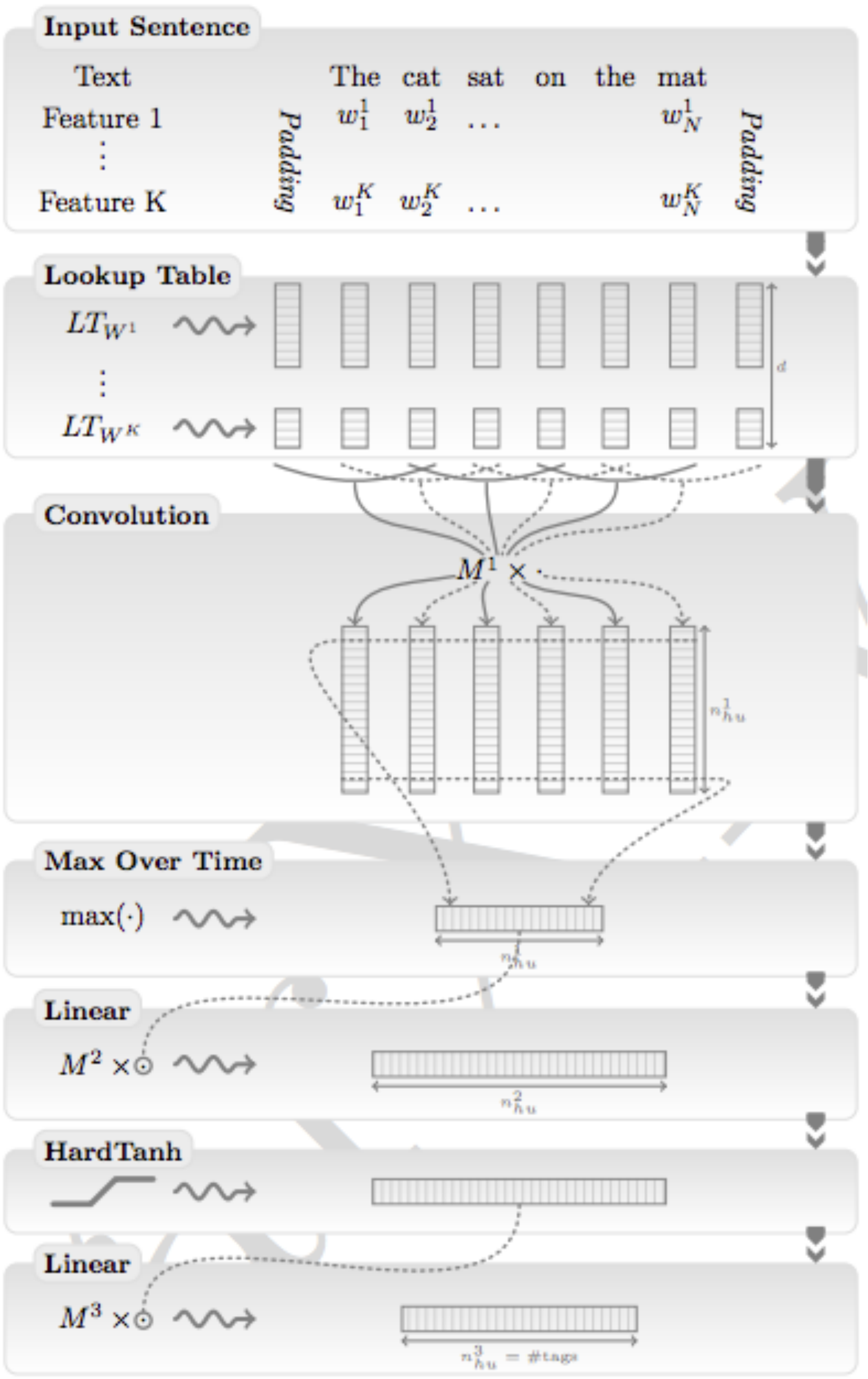
(1,0,0,1,0,1,1 .....,0,1)

分类器

正面评价　　负面评价

词是自然语言处理的基础

- 神经网络初始化



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]

- 问题：
  - 语义鸿沟问题
  - Cosine（减肥，瘦身）= 0
  - 维数灾难、稀疏
  - 无法表示unseen words

词表

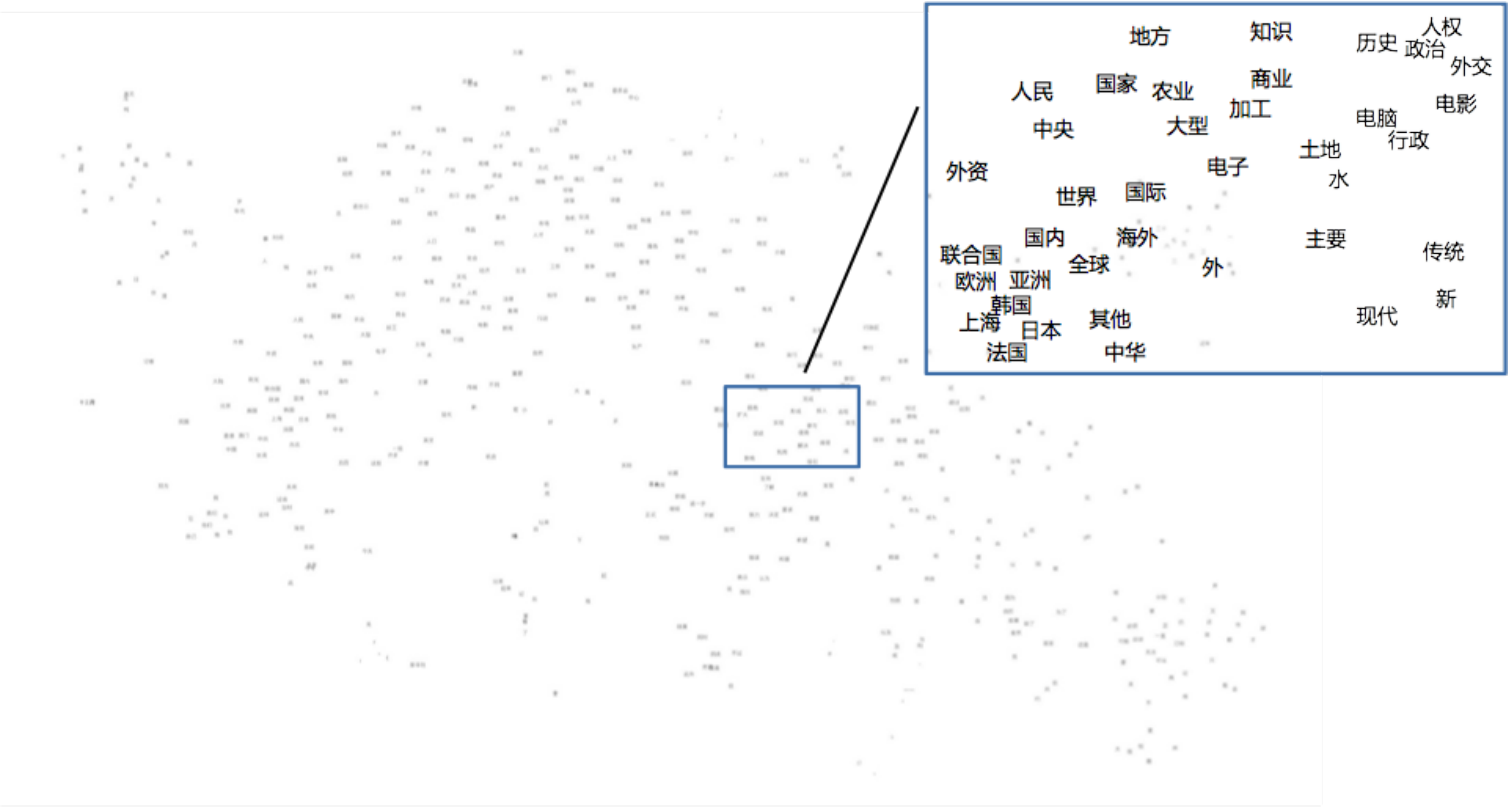
瘦身
人民
国家
减肥
成都
北京
美国
中科院
机器
学习
的
.
.
.

词表示

- 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]

- 每一维可以看成词的语义或者主题信息
- 维度压缩
- 很好的解决语义鸿沟问题
- Cosine（减肥，瘦身）= 0.7635
- 基于学习模型，可以快速对于unseen words进行表示

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
中央
大型
电脑
行政
土地
外资
电子
水
世界
国际
联合国
国内
海外
主要
传统
全球
外
欧洲
亚洲
新
韩国
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]

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

the doctor. </p><p> `Just checking on the **bardiwac** ,' he boomed as he came back. `Edith's very
</p><p> `I hope you'll take to a good French **bardiwac** ,' chimed in Arthur Iverson jovially. `One
`Our host did slip out to attend to the **bardiwac** …' </p><p> `That was before the shrimp
Iverson did when he went through to see to the **bardiwac** before dinner.' Henry rubbed his hands.
and drinking red wine from France -- sour **bardiwac** , which had proved hard to sell. The room
eyes were alight and he was drinking the **bardiwac** down like water. `It is like Hallow-fair
quizzically at him and offering him some more **bardiwac** . </p><p> He shook his head. `I will sleep
drinks (as Queen Victoria reputedly did with **bardiwac** and malt whisky), but still the result
Do we really `wash down' a good meal with **bardiwac** ? Port is immediately suggested by Stilton
completely different: cheap and cheerful **bardiwac** . Two good examples from Victoria Wine are
examples from Victoria Wine are its house **bardiwac** , juicy and a touch almondy, a good buy
opened a bottle of rather rust-coloured **bardiwac** . I ate too much and drank nearly three-quarters
elections, it was apparent the SDP of` **bardiwac** and chips' mould-breaking fame at the time
the black hills. Not a night of vintage **bardiwac** . </p><p> Burnley: Pearce, Measham, McGrory
SONS Old School -- the Marlborian navy, **bardiwac** and slim-white stripe. Heavy woven silk
white-hot passion. We are like a good bottle of **bardiwac** ; we both have sediment in our shoes. </p>
few minutes later he was uncorking a fine **bardiwac** in Masha's room, saying he had something
the phone. Surkov silently offered me more **bardiwac** but I indicated a bottle of Perrier. </p>
defenders as Villa swept past them like a **bardiwac** and blue tidal wave. </p><p> Things are difficult
campaign. Refreshed by a nimble in-flight **bardiwac** , they serenaded him with a special song

	glass	drink	grape	red	meal
bardiwac	10	22	43	16	29
car	5	0	0	10	0

Paradigmatic vs. Syntagmatic

Syntagmatic Relation

You like machine learning

Paradigmatic Relation

We like deep learning

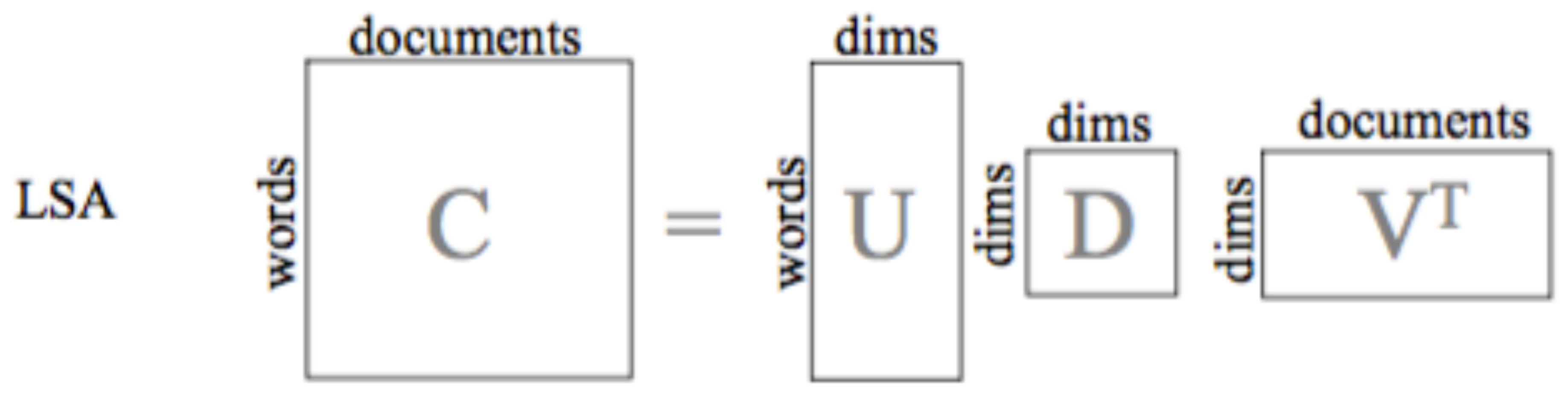
Syntagmatic Relation

传统词向量学习方法

- “词-文档”共现矩阵
  - Latent Semantic Analysis (LSA)
  - Probabilistic Semantic Analysis (PLSA)

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

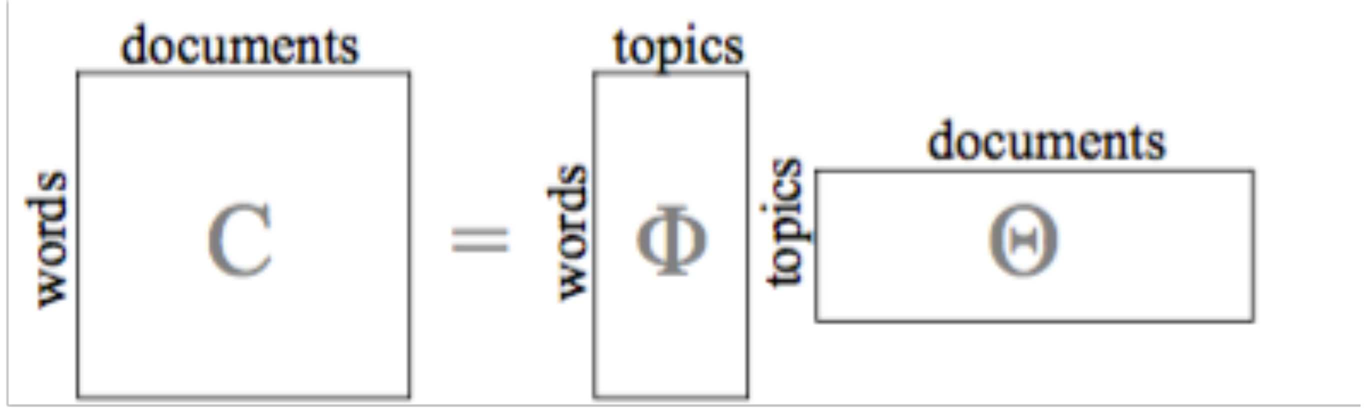
Latent Semantic Analysis (LSA)



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

$$\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$$

Topic Model



human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

传统词向量方法

- “词-词”共现矩阵
  - Brown Clustering [Brown et al. 1992]
  - Hyperspace Analogue to Language, HAL [Lund et al. 1996]
  - GloVe [Pennington et al 2014]

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

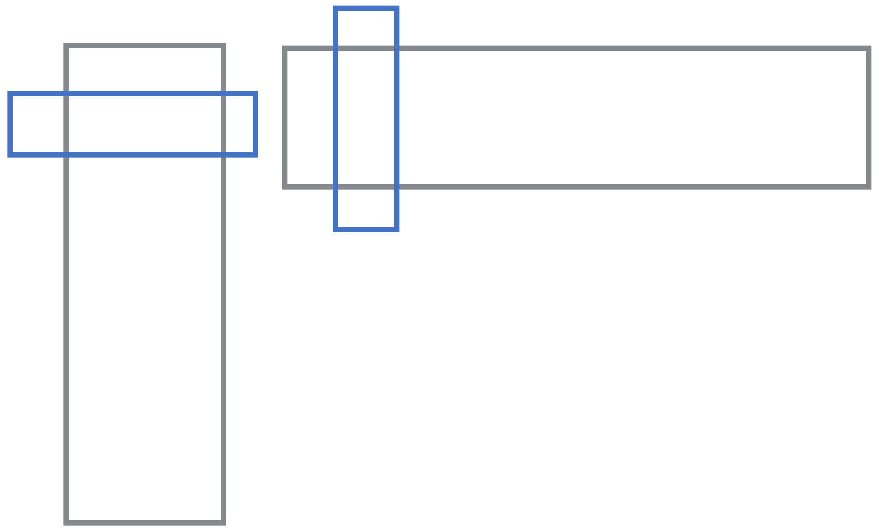
	w1	w2	w3	w4
w1		2	4	1
w2	2		3	
w3	4	3		1
w4	1		1	

GloVe

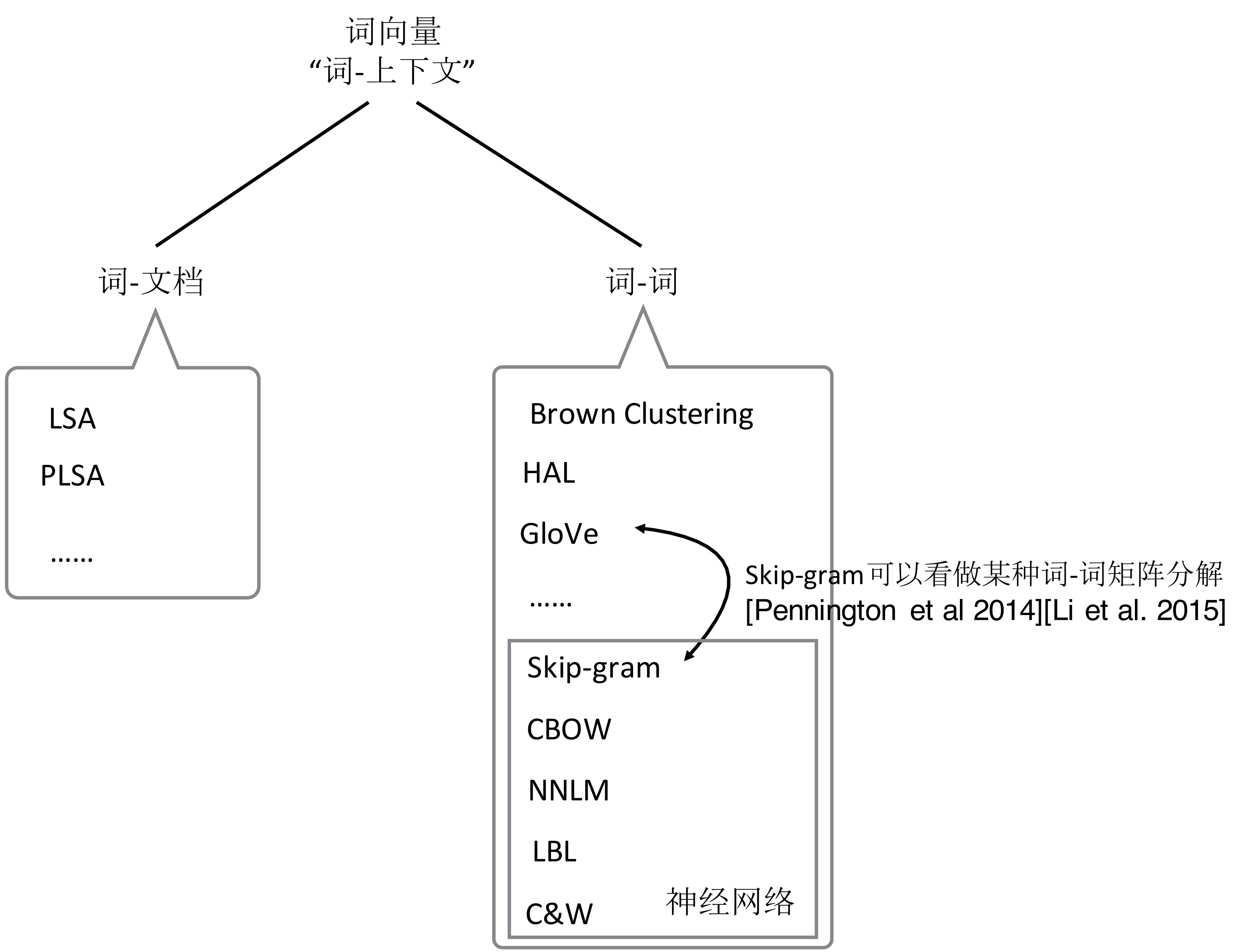
$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

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	



RoadMap



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

如何通过神经网络的方法训练得到一组词向量？
如何训练得到一组好的词向量？

语言模型

- 目标：计算一个词串的概率

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

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

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

例子

他是研究生物的

他 是 研究生 物 的

他 是 研究 生物 的

p(Seg1) =p(他|<BOS>)×p(是|他)×p(研究生|是)×p(物|研究生)×p(的|物)×p(的|<EOS>)

p(Seg2) = p(他|<BOS>)×p(是|他)×p(研究|是)×p(生物|研究)×p(的|生物)×p(的|<EOS>)

语言模型

Representation

Classifier

$$p(quick|C) \approx p(fast|C)$$

$$R(fast) \approx R(quick)$$

NNLM

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

$$P(w_1,\dots,w_N)=\prod_t P(w_t \mid w_{t-1},\dots,w_{t-n+1})$$

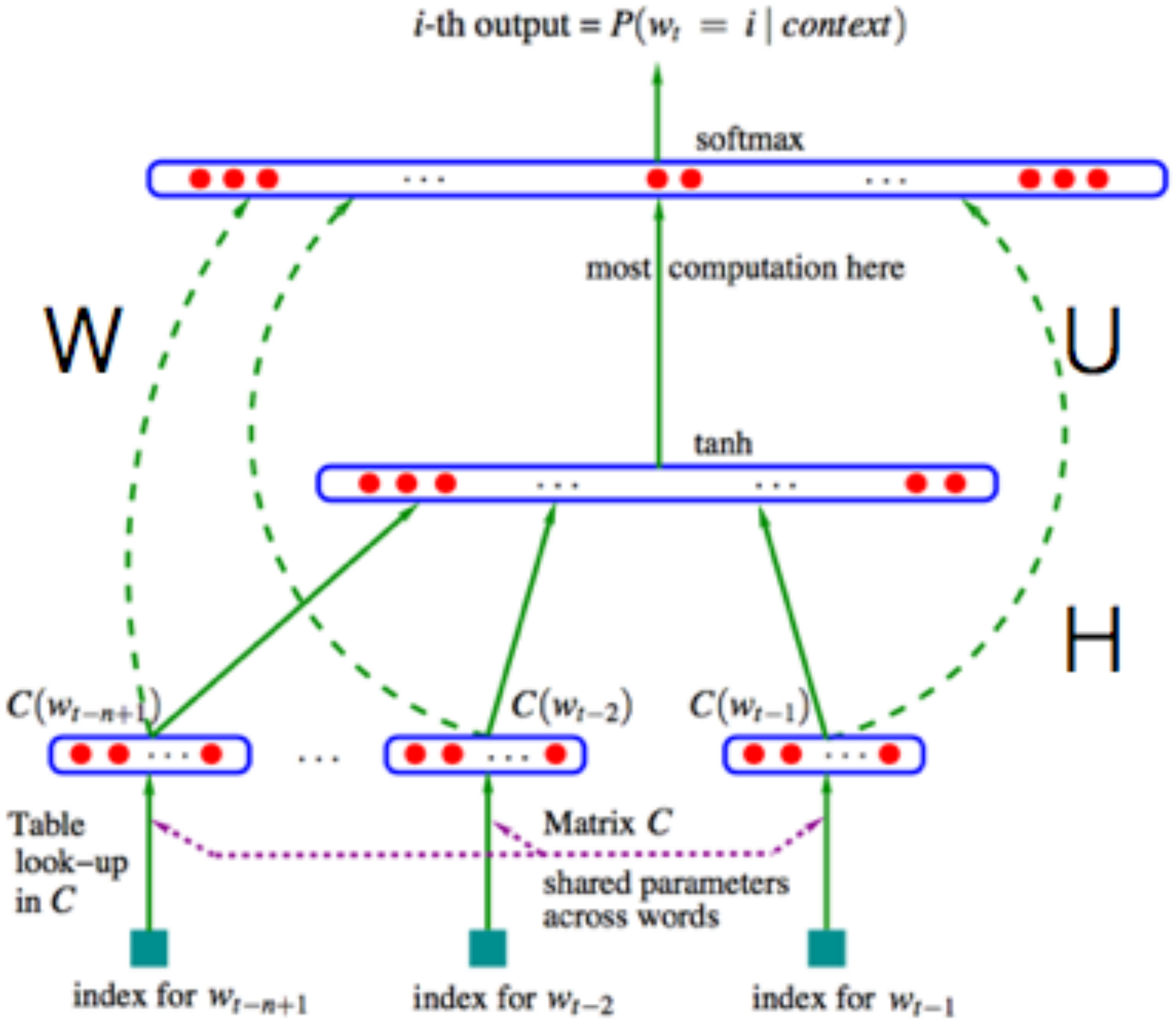
$$f(w_t,w_{t-1},\cdots,w_{t-n+1})=\hat{P}(w_t \mid w_{t-1},\cdots,w_{t-n+1})$$

$$f(w_t,w_{t-1}\dots,w_{t-n+1})=g(w_t,C(w_{t-1})\dots,C(w_{t-n+1}))$$

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

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

NNLM



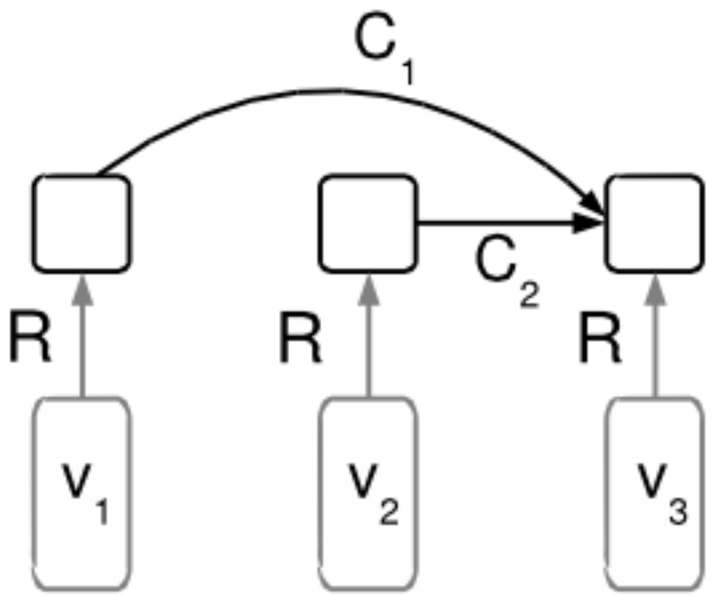
$$y = b + Wx + U\tanh(d + Hx)$$

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

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

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

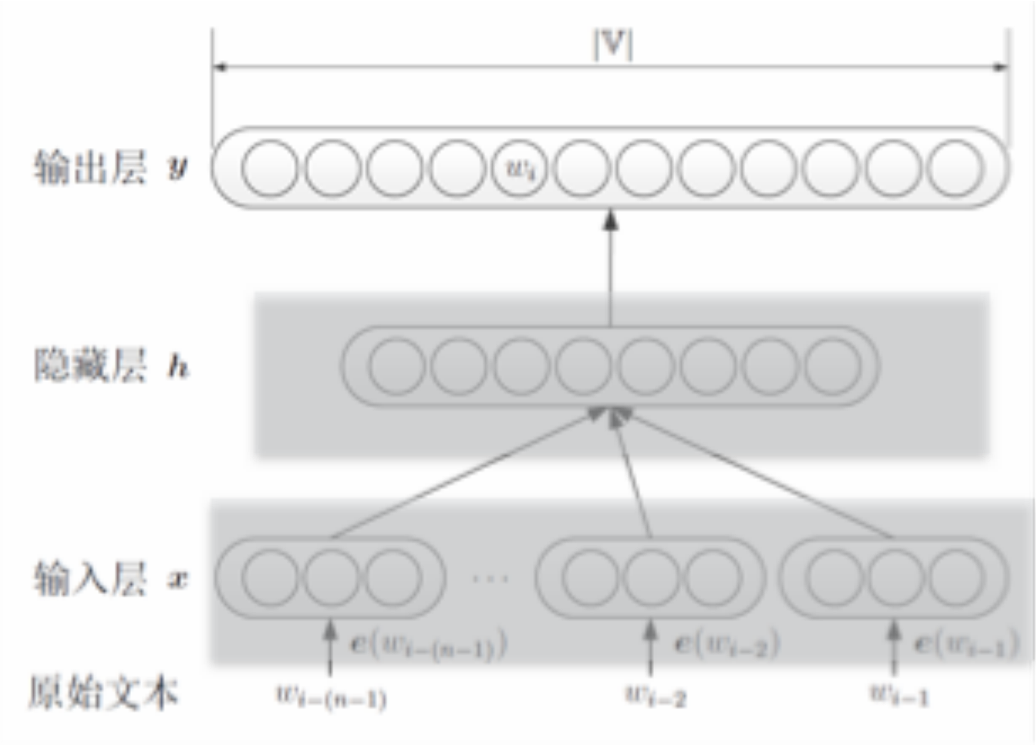
$$E(w_n;w_{1:n-1}) = -\left(\sum_{i=1}^{n-1} v_i^T R C_i\right) R^T v_n - b_r^T R^T v_n - b_v^T v_n.$$

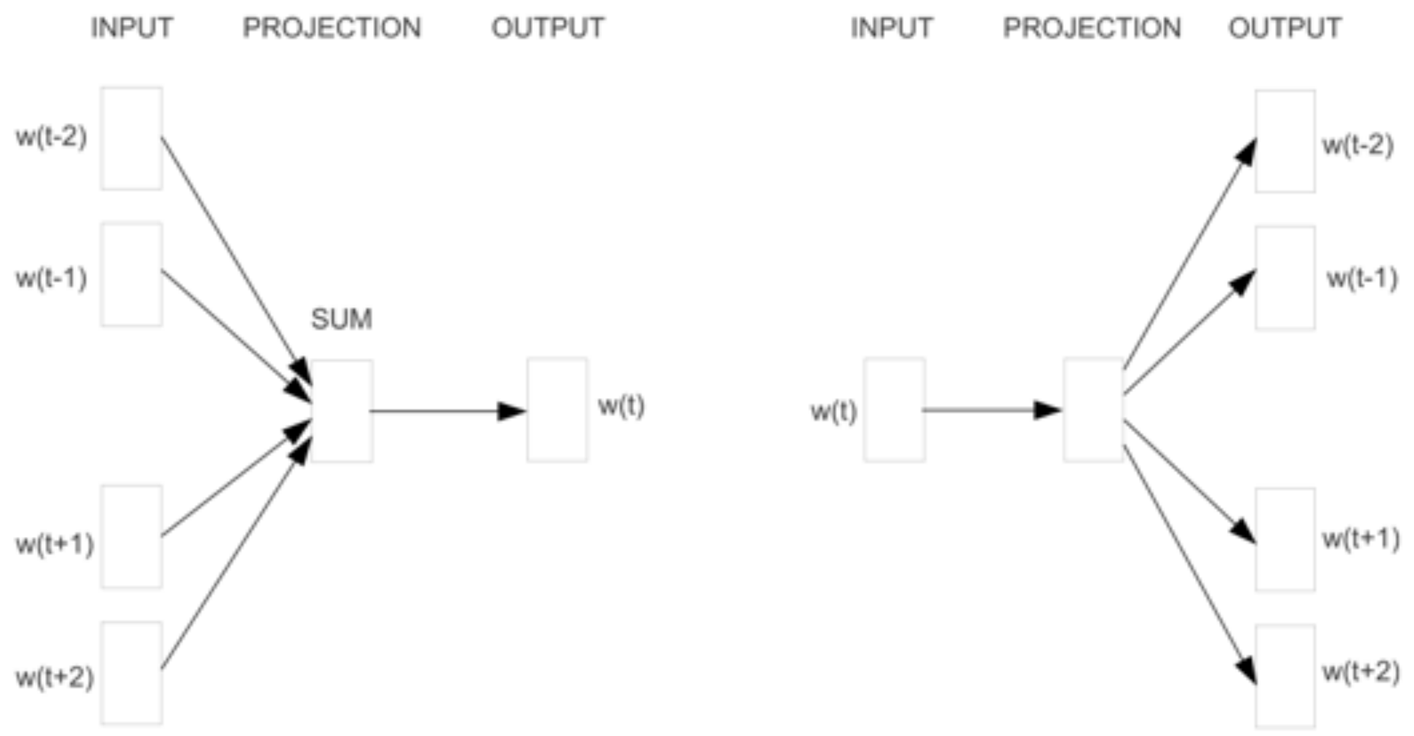
$$Z_c = \sum_{w_n} \exp(-E(w_n;w_{1:n-1}))$$

CBOW / Skip-gram

- 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

[T. Mikolov et al, ICLR 2013]

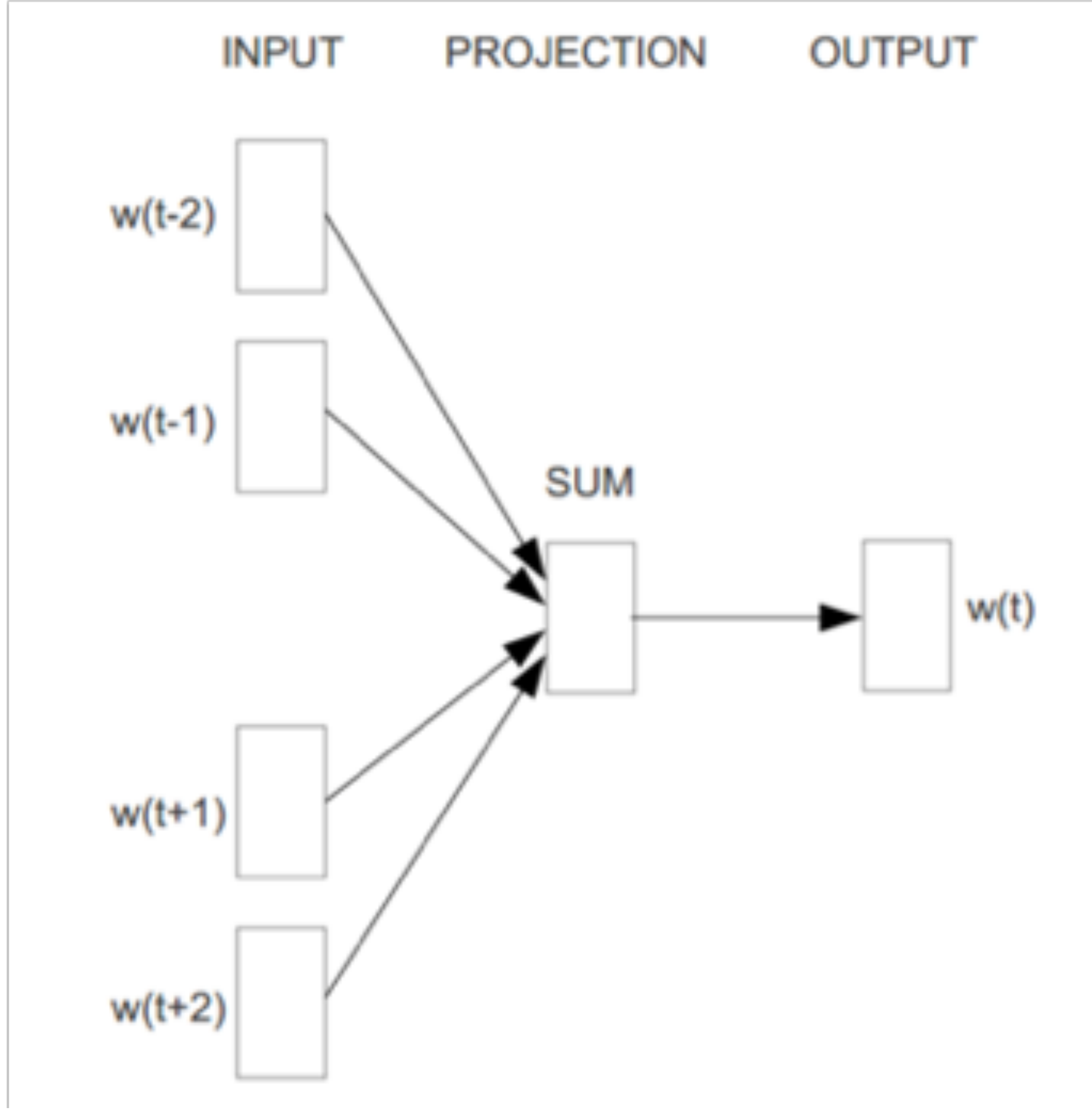
CBOW

- Continued Bag of Words Model

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

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

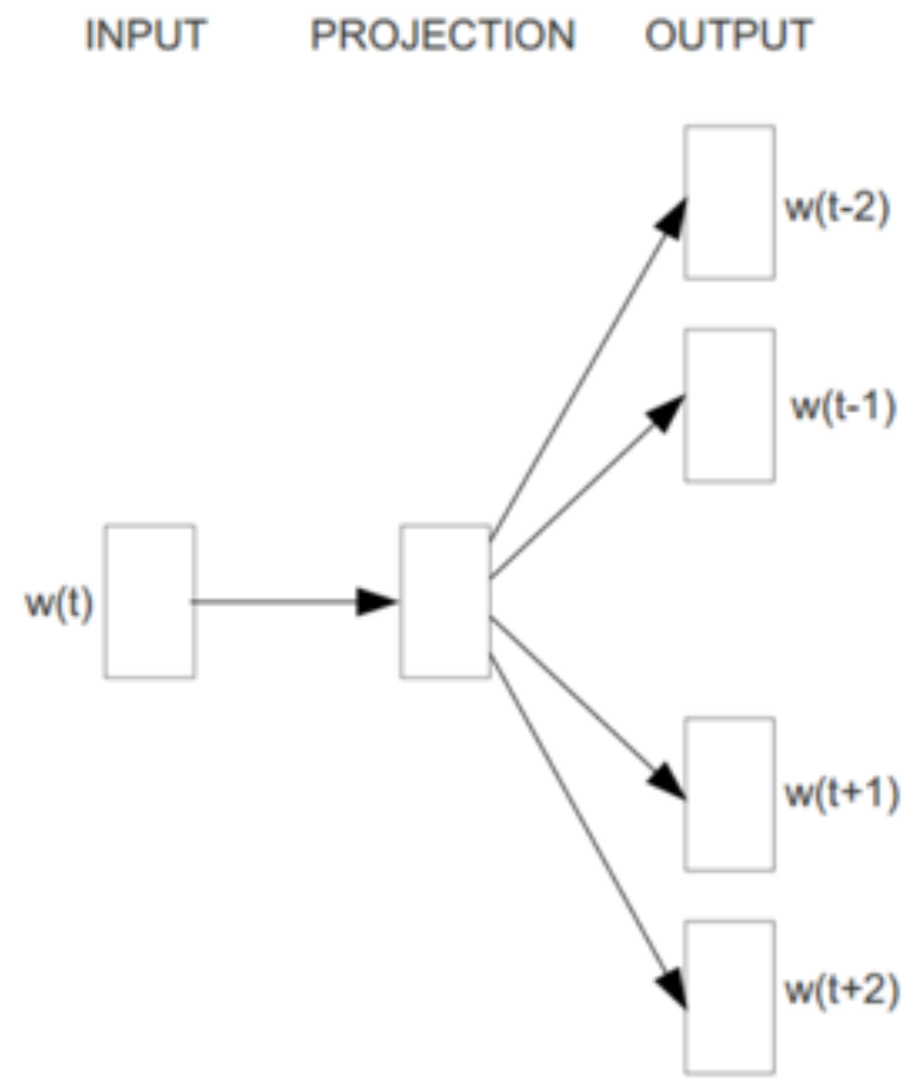
$$v_{C_i} = \sum_{j \in C_i} v_j$$



Skip-Gram

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

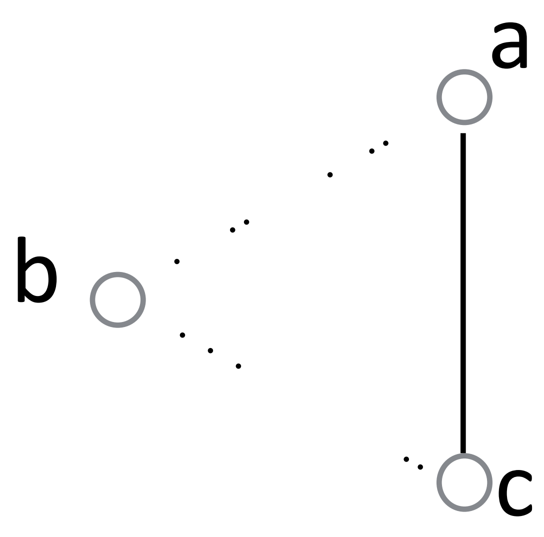
$$P(w_i \mid w_j) = \frac{\exp(v_i'^{T} v_j)}{\sum_{w_i}\exp(v_i'^{T} v_j)}$$



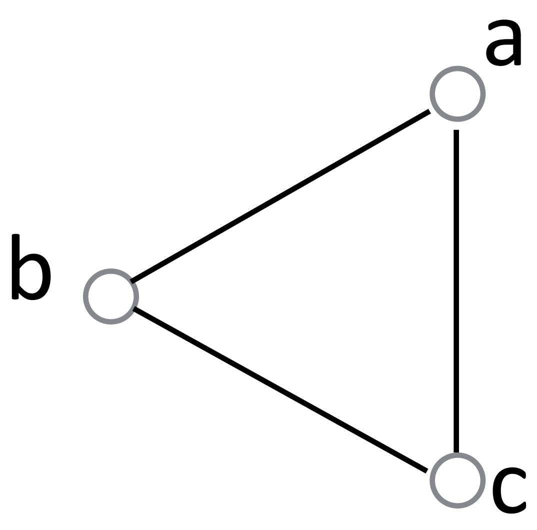
Contextual Vector

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

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



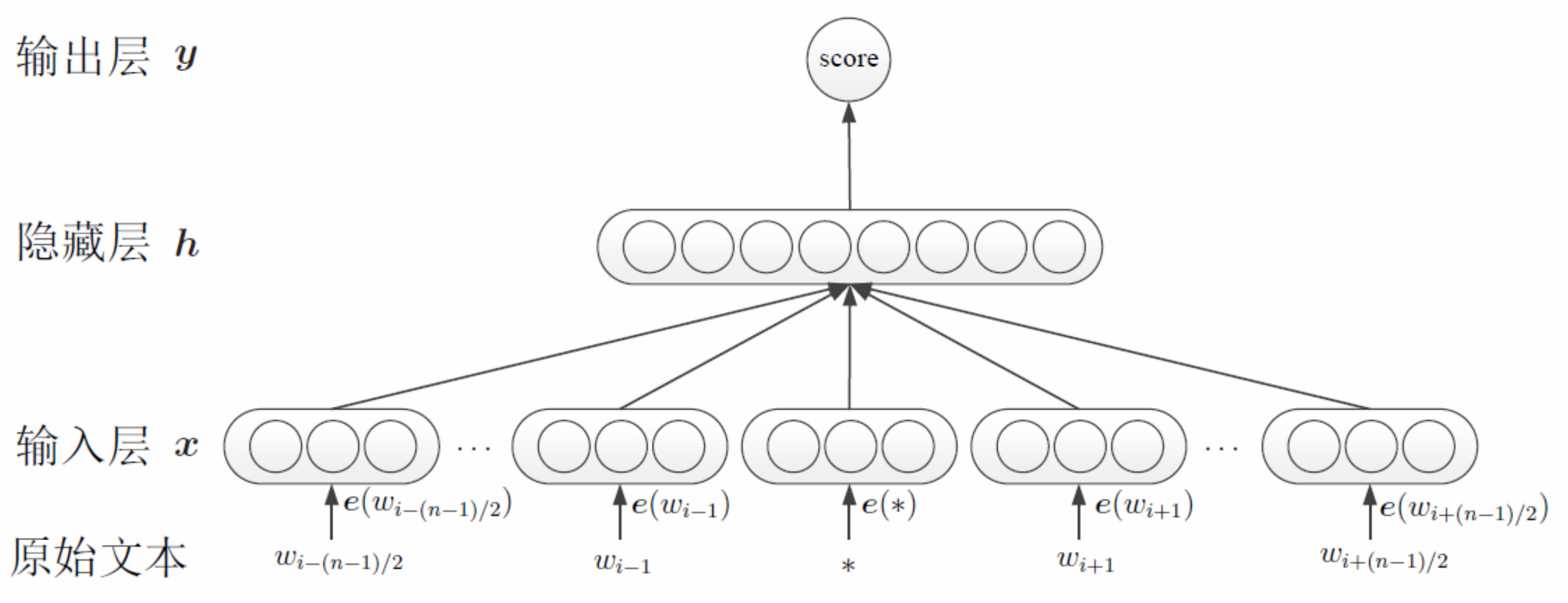
Paradigmatic Relation



Syntagmatic Relation

C&W

- 目标：词向量



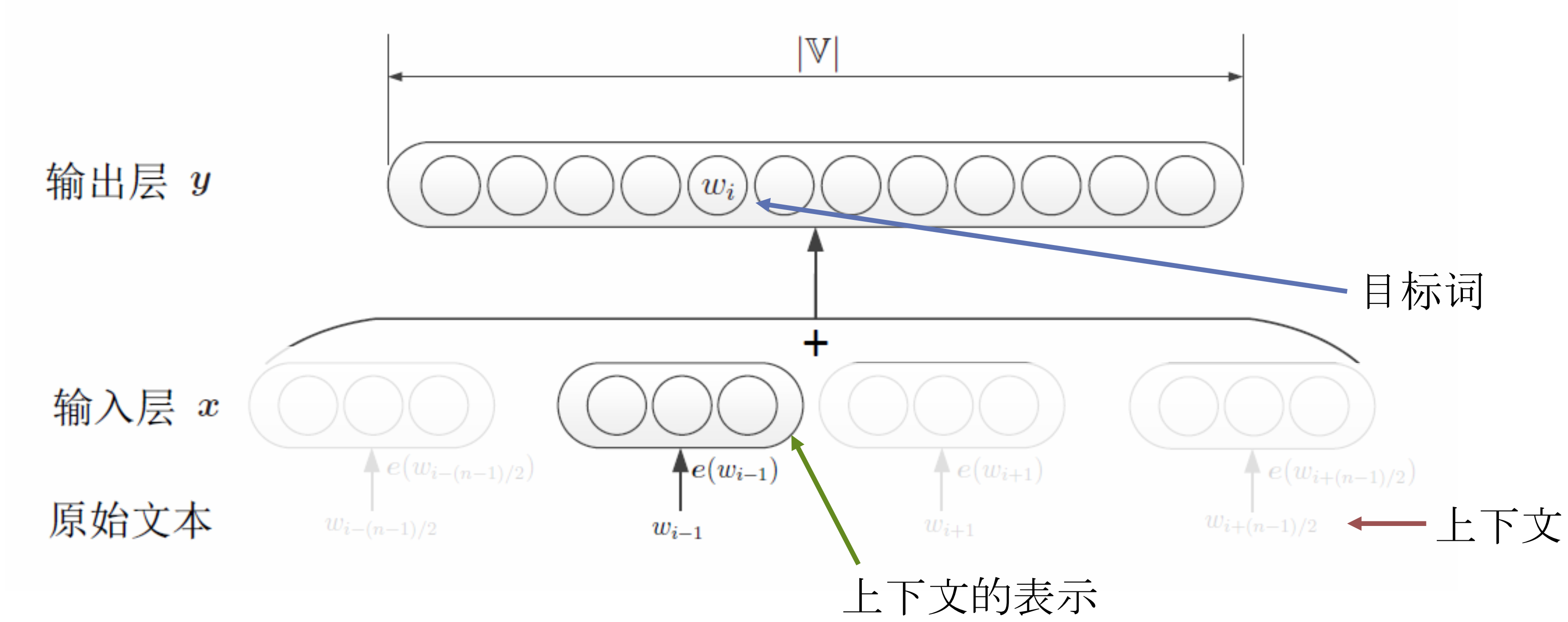
目标函数 $\max(0, 1 - s(w, c) + s(w', c))$

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

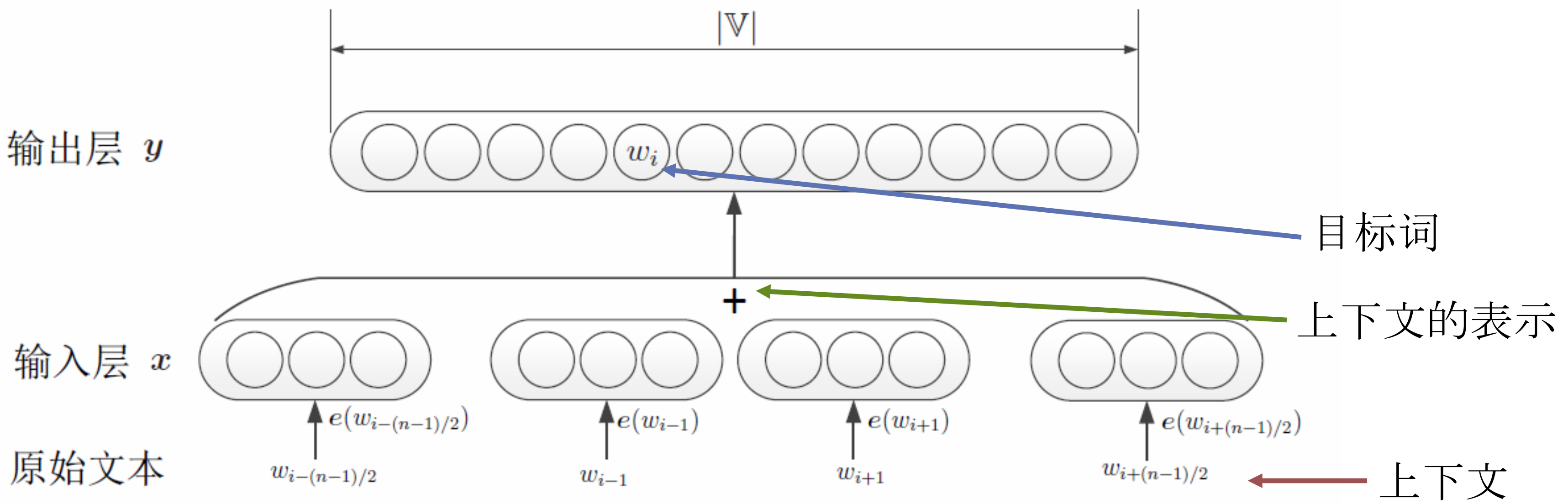
Skip-gram



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



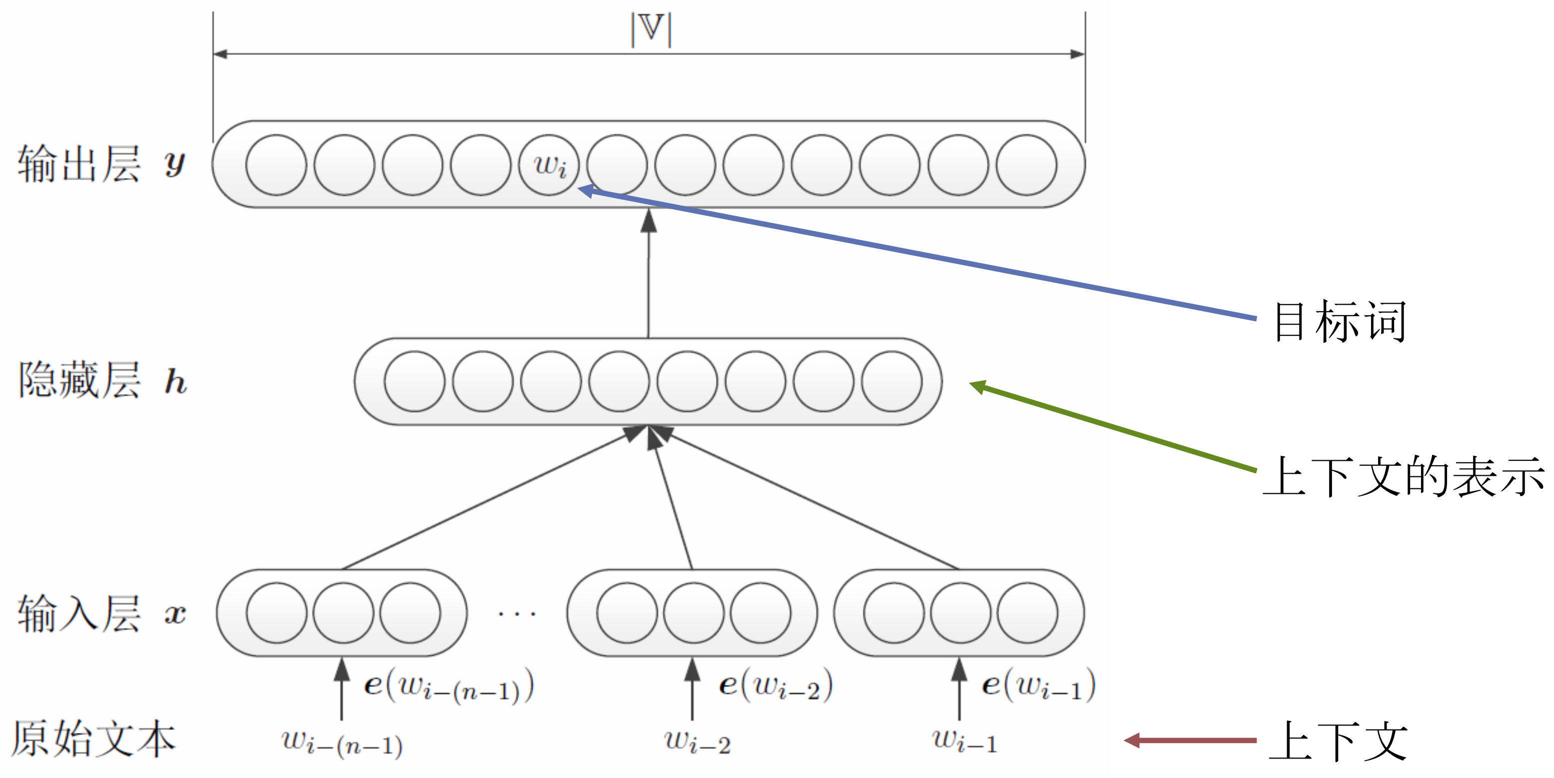
Continuous Bag-of-Words

目标词和上下文的关系： $P(w_i | C_i)$

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

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

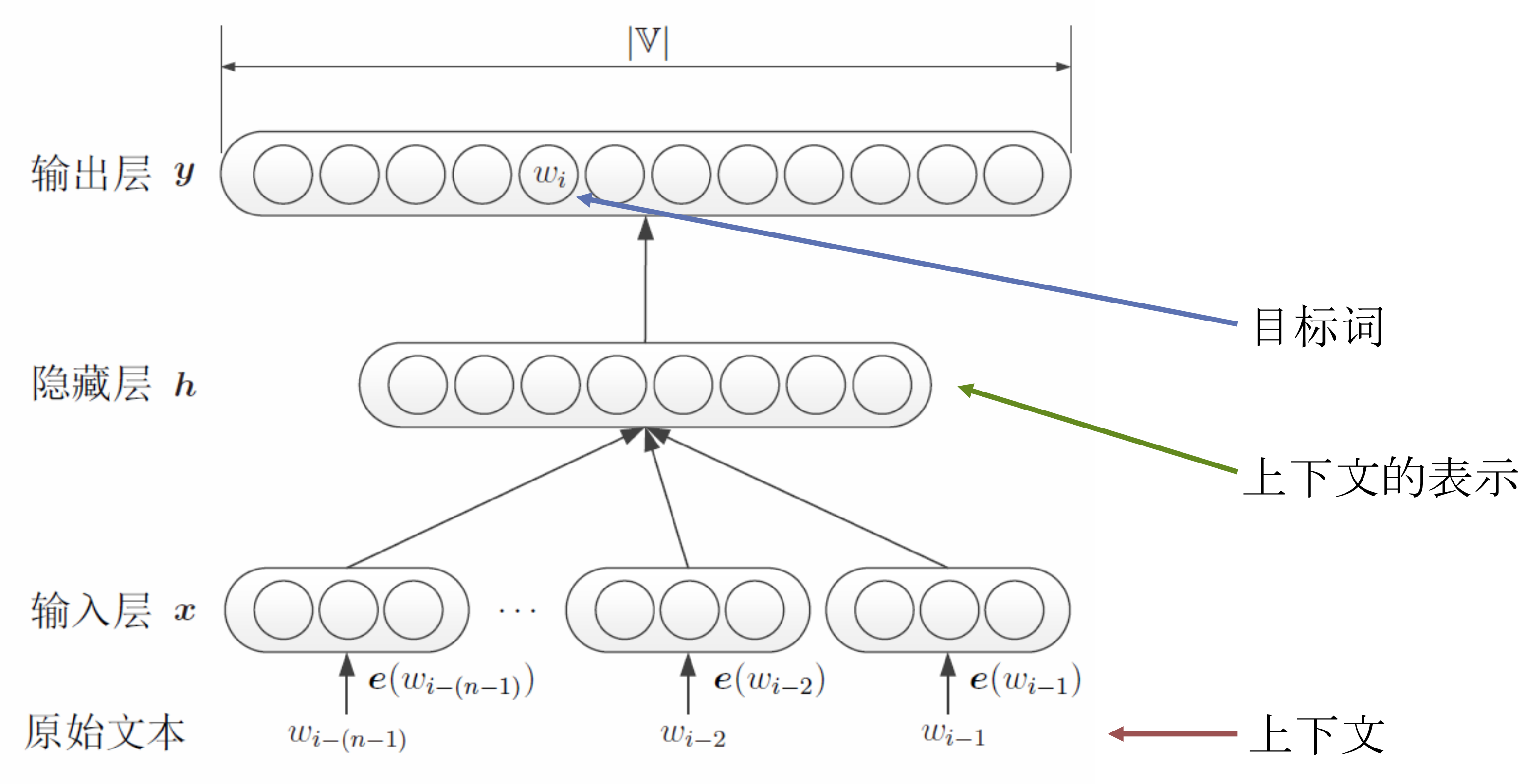
LBL



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

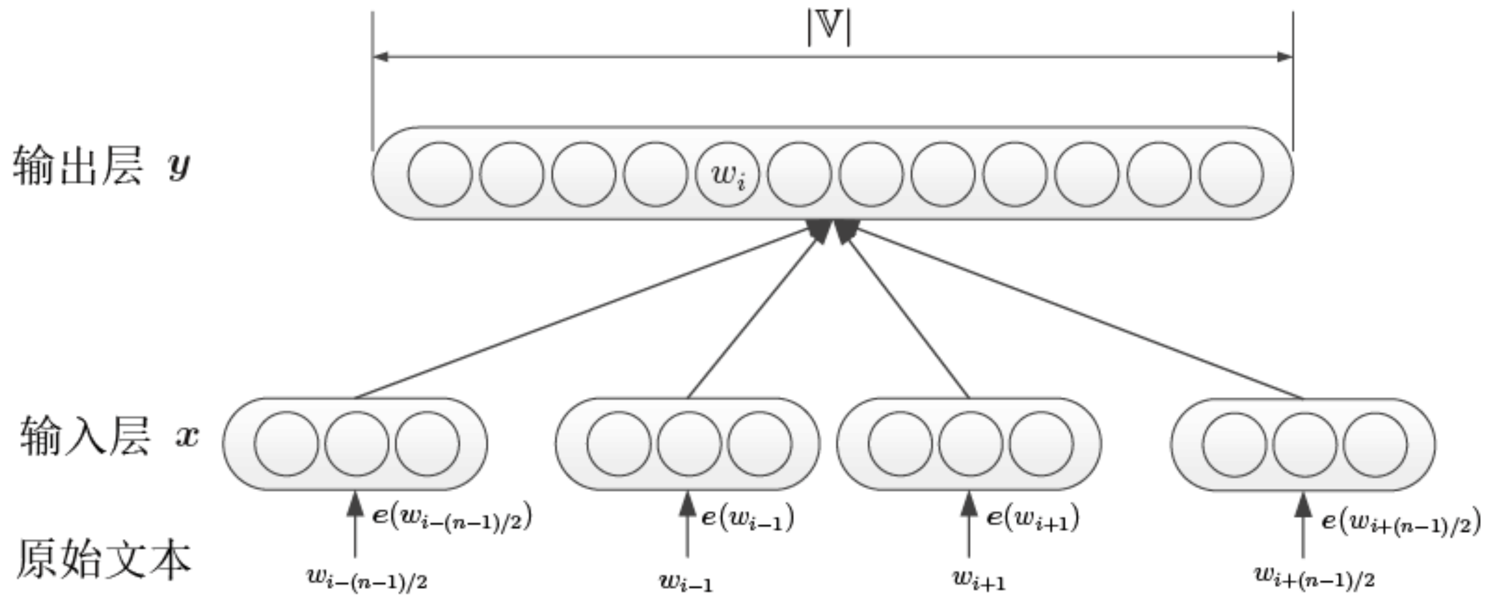
NNLM



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

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

Order(Virtual Model)

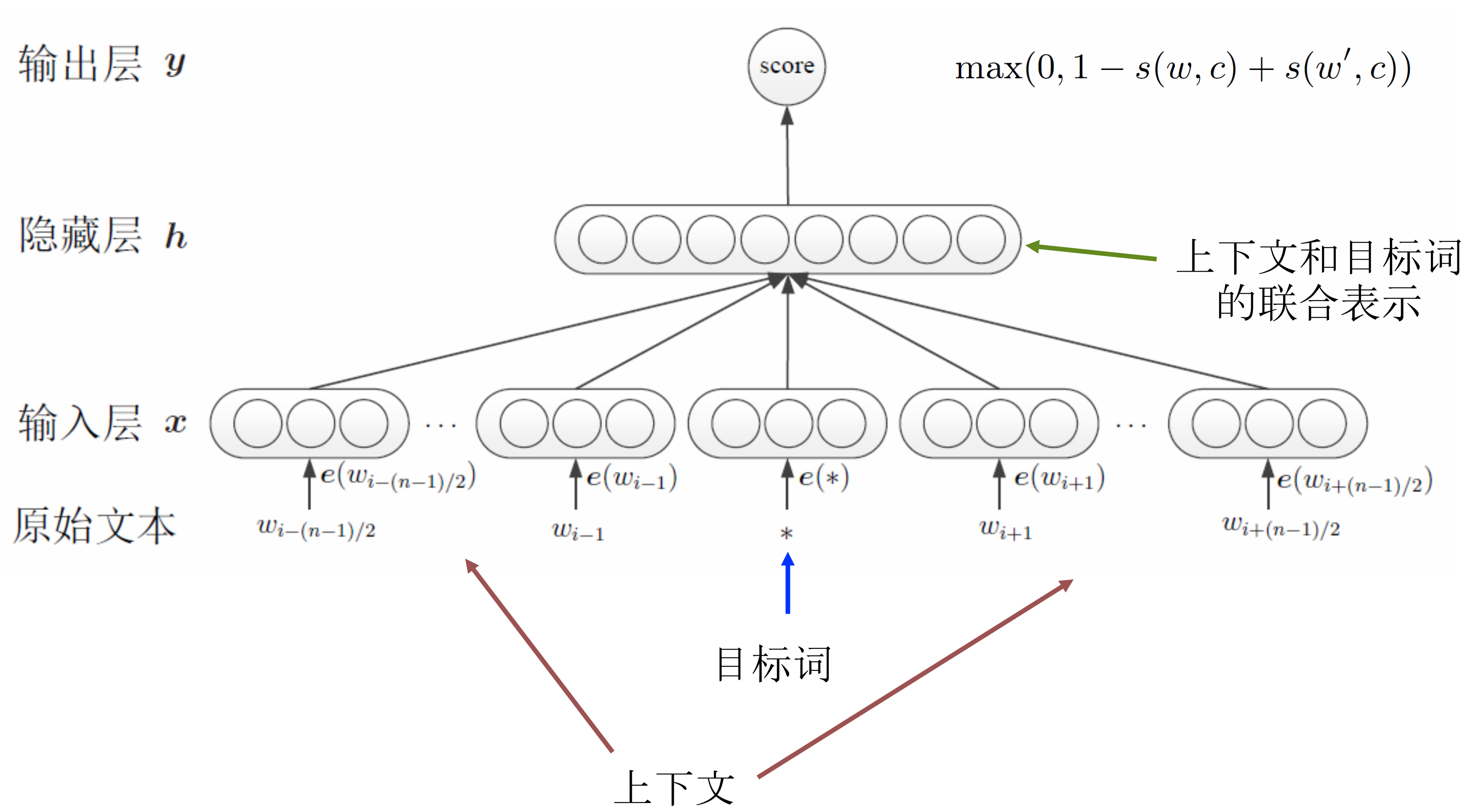


目标词和上下文的关系： $P(w_i | C_i)$

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

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	c predicts w	one of c
CBOW	c predicts w	average of c
Order	c predicts w	concatenation
LBL	c predicts w	compositionality
NNLM	c predicts w	compositionality
C&W	scores w, c	compositionality

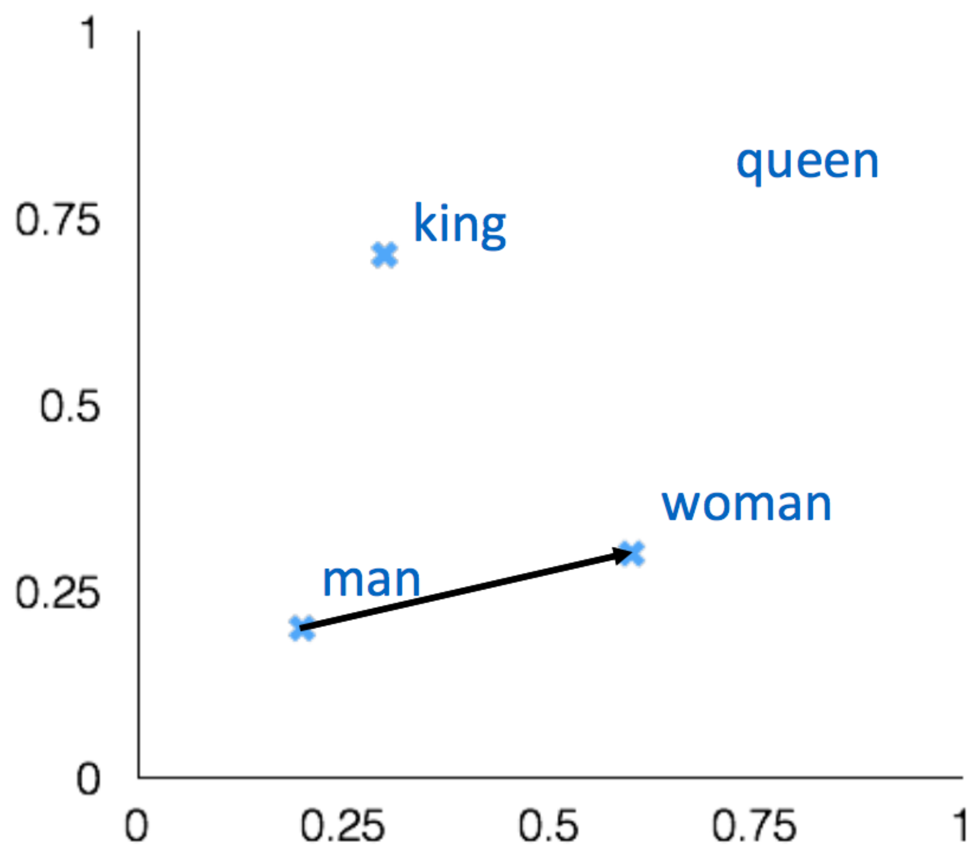
简单

复杂

怎样才算是好的词向量

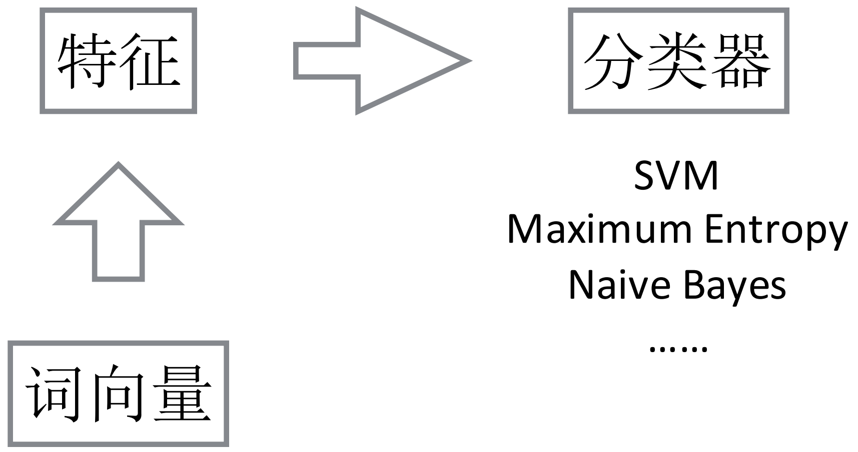
词向量应用

- 语言学应用



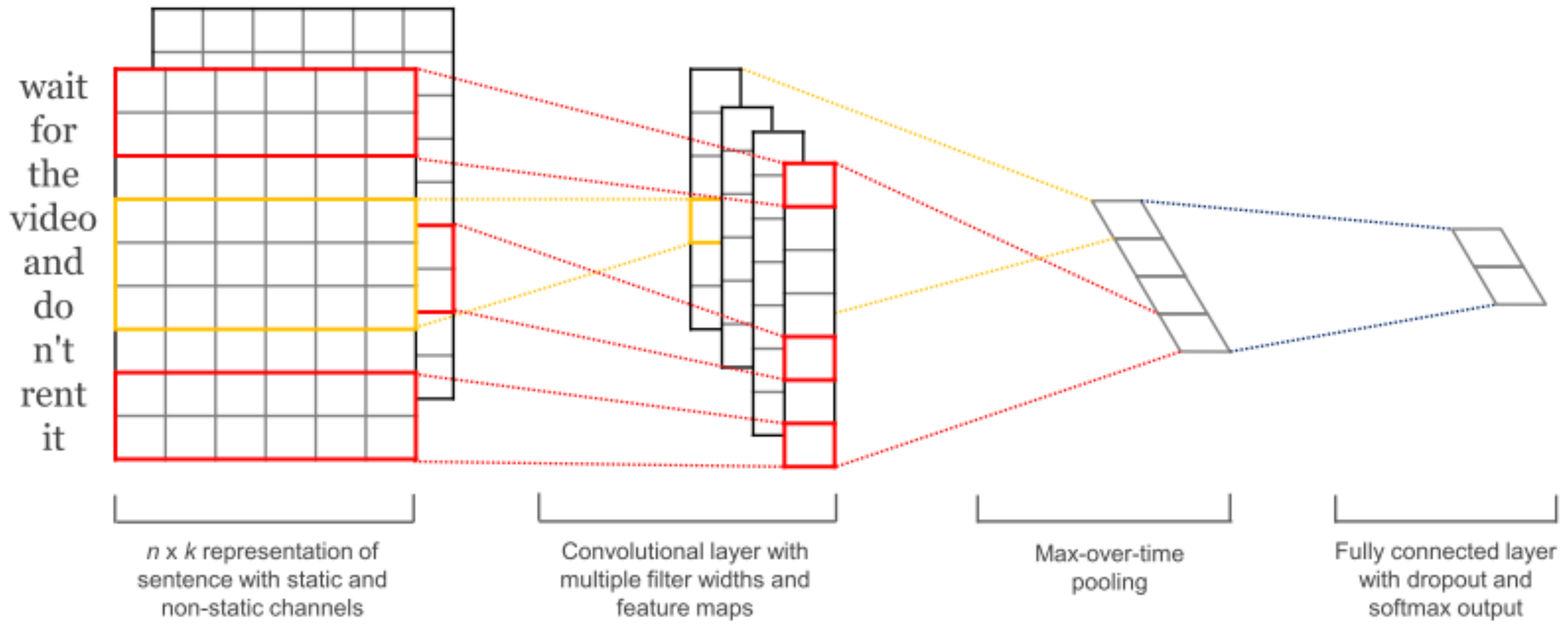
词向量应用

- 作为某一任务的特征
  - 文本分类
  - 情感分类
  - 传统特征：unigram、bigram、trigram
  - 分布式特征：Word Embeddings



词向量应用

- 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
  - predict – predicting ≈ dance – dancing
- 类比关系（语义）（sem）9k
  - king – 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

- 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

[L. Finkelstein et al., 2013]

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

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

- 任务：找给定词语的同义词（tfl）80个选择题 [T. Landauer & S. Dumais, 2013]

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

- 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

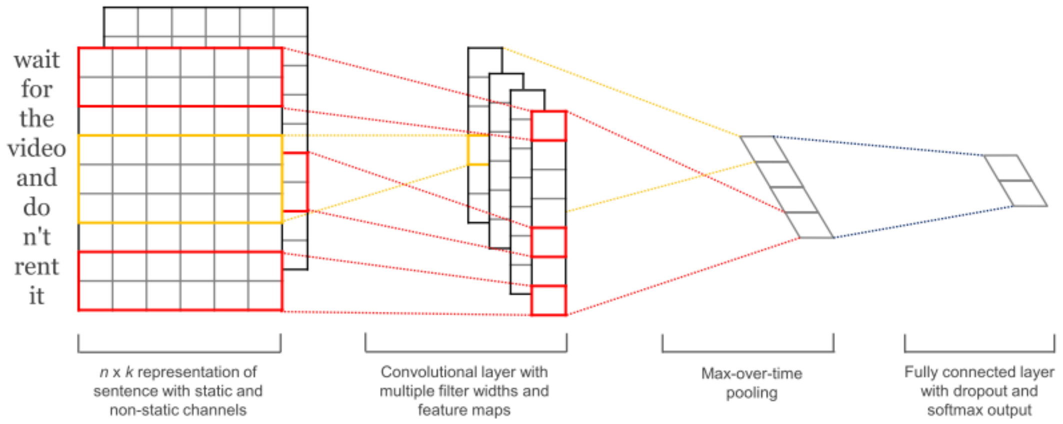
[Turian et al., 2010]

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

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

[Y. Kim, 2014]

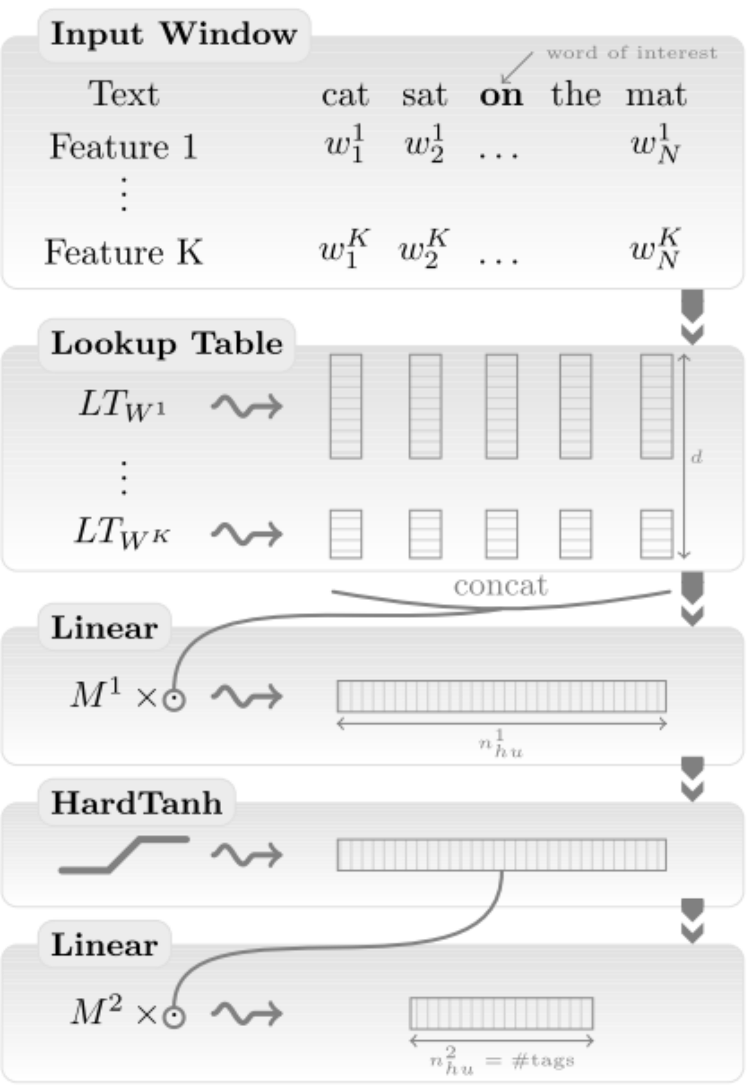


Model	cnn
Random	36.60
Skip-gram	43.84
CBOW	43.75
Order	**44.77**
LBL	43.98
NNLM	44.40
C&W	41.86

评价任务：词性标注

- 任务：标注给定句子中词的词性（pos）数据规模 [R. Collobert et al., 2011]
- 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

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

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

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

复杂

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

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，相对复杂的模型展现较好的结果

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	$\boxed{93}$	52	90	98	50	76	85	96
100M	$\boxed{76}$	30	88	93	46	77	83	86
Wiki 1.6B	$\boxed{92}$	**100**	**100**	93	51	**100**	86	94
100M	$\boxed{74}$	65	98	93	47	88	90	83
W&N 2.8B	$\boxed{\mathbf{100}}$	89	95	93	50	97	91	**100**
1B	$\boxed{98}$	87	95	**100**	48	98	90	98
100M	$\boxed{79}$	63	97	96	51	85	92	86
10M	$\boxed{29}$	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），维基百科语料具有优势

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

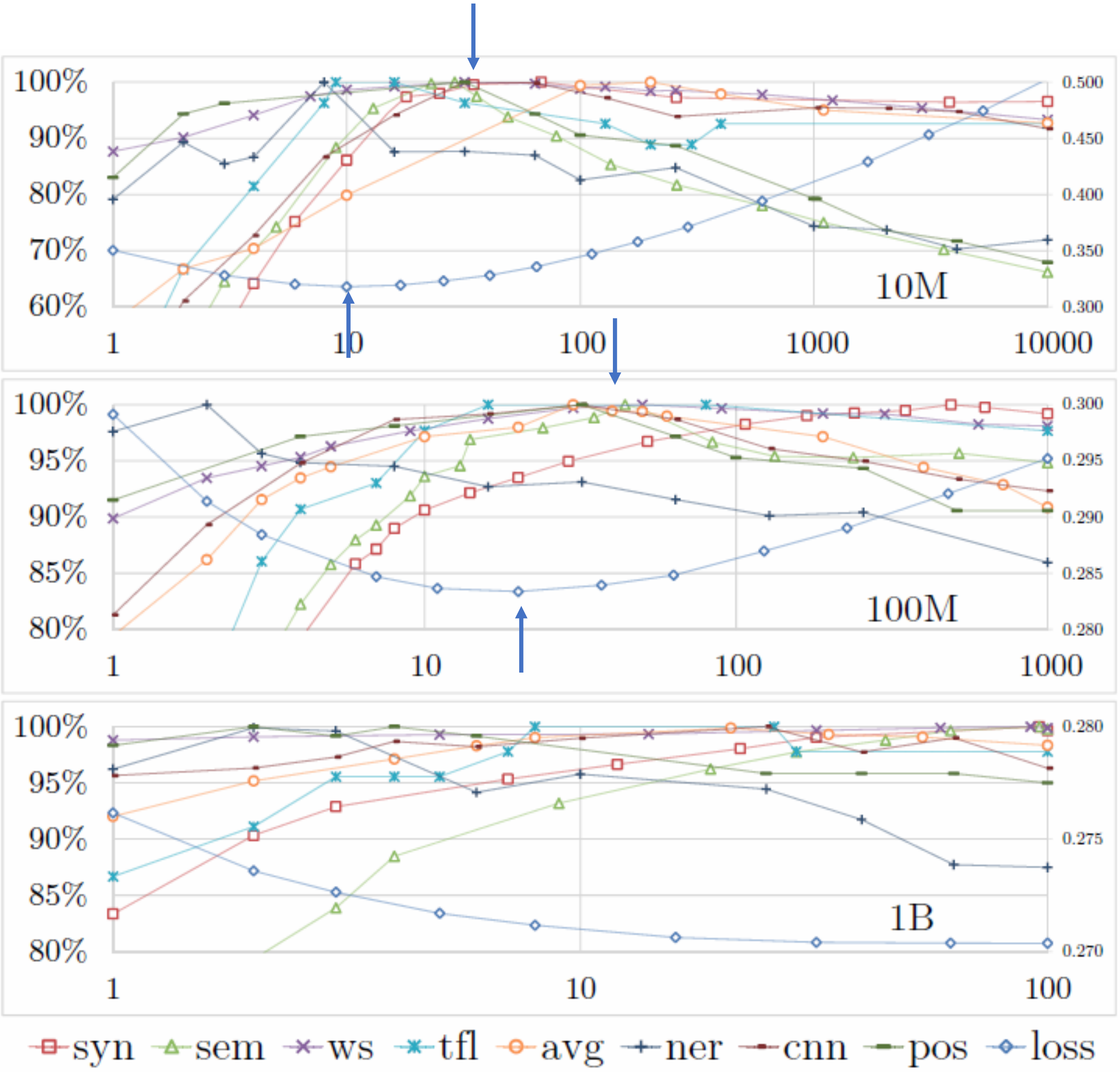
W&N \ IMDB	20%	40%	60%	80%	100%
+0%	91	94	100	100	100
+20%	79	87	91	96	99
+40%	68	86	88	92	98
+60%	65	79	85	88	93
+80%	64	75	84	87	92
+100%	64	70	83	86	88

CBOW

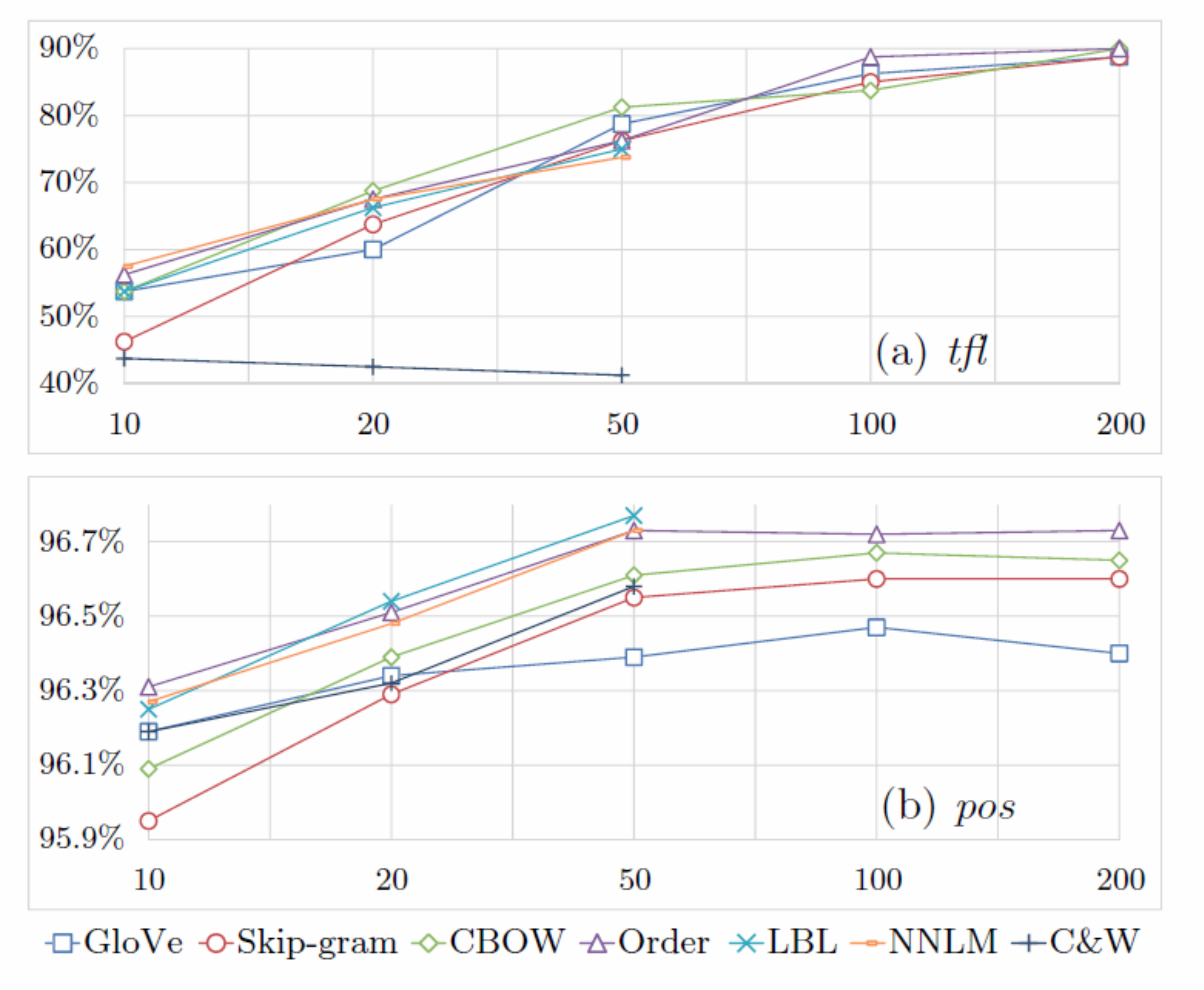
领域更加重要

训练参数：Iteration Number

- Early stop



训练参数：Dimension



Taking Home Message

- 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

Thanks