

# 第7章 文本表示


宗成庆

中国科学院自动化研究所

[cqzong@nlpr.ia.ac.cn](mailto:cqzong@nlpr.ia.ac.cn)

# 本章内容

---

-  1. 问题提出
- 2. 向量空间模型
- 3. 表示学习模型
- 4. 习题

# 1. 问题提出

## ◆文本

文本是由文字和标点组成的字符串，词汇、短语、句子、段落和篇章都是不同粒度的文本。

## ◆文本表示的目的

用形式化的方法描述文本，既能反映文本的内容，也能体现不同文本之间的差异，且可计算。

# 本章内容

---

1. 问题提出

 2. 向量空间模型

3. 表示学习模型

4. 习题

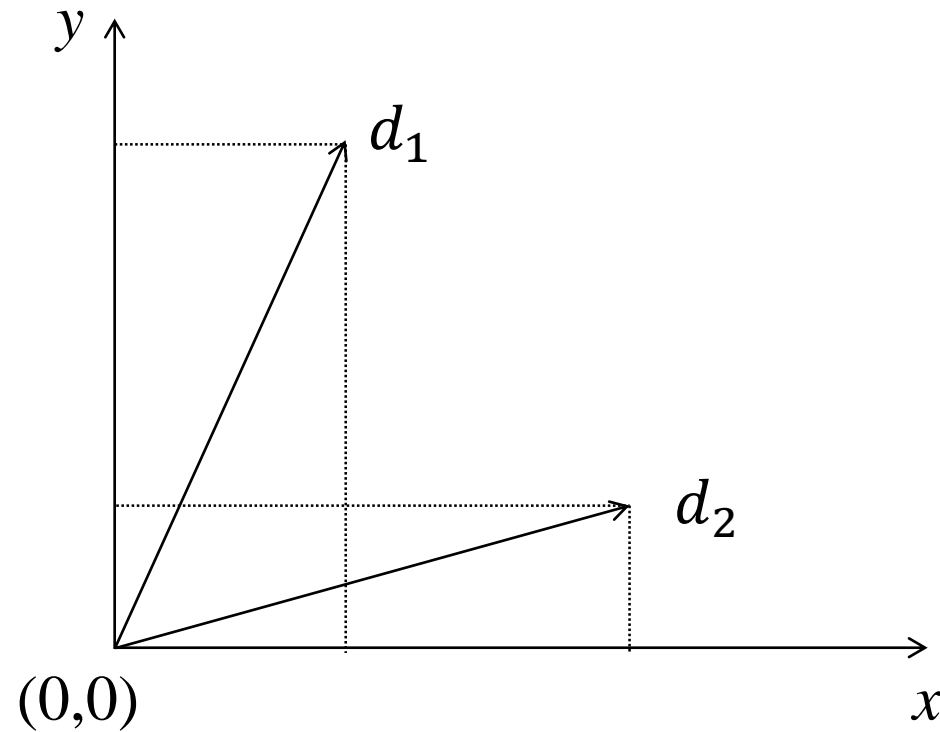
## 2. 向量空间模型

### ◆ 概念解释

- 向量空间模型(vector space model, VSM) 由G. Salton等人于1960s末期在信息检索领域中提出，核心思想是将文本视为特征项的集合。
- **特征项：** VSM中最小的语言单元，可以是字、词、短语等。文本表示为特征项集合 $(t_1, t_2, \dots, t_n)$
- **特征项权重：** 每个特征项在文本中的重要性不尽相同，用 $w$ 表示特征项 $t$ 的权重，相应地，文本可以表示为 $(t_1:w_1, t_2:w_2, \dots, t_n:w_n)$ ，或 $(w_1, w_2, \dots, w_n)$

## 2. 向量空间模型

文本1:  $d_1 = (w_1, w_2, \dots, w_n)$       文本2:  $d_2 = (w'_1, w'_2, \dots, w'_n)$



## 2. 向量空间模型

### ➤ 特征项

- 单词、词组/短语或 $n$ -gram: 若单词作为特征项, 那么特征项的集合可视为一个词汇表。
- 特征项可以从语料中统计获得。所有的特征项可看作一个“词”袋(集合), 所以向量空间模型被称为词袋模型(bag-of-words, BOW)。



## 2. 向量空间模型

### ➤ 特征项的权重

- 表示每个特征的重要性。

✓ 布尔变量（是否出现）： $w_i = \begin{cases} 1, & \text{如果 } w_i \text{ 在文本 } d \text{ 中} \\ 0, & \text{否则} \end{cases}$

✓ 项频（term frequency, TF）： $w_i = \log(tf_i + 1)$

文档 特征项	$d_1$	$d_2$
$t_1$	$tf_{11}$	$tf_{12}$
$t_2$	$tf_{21}$	$tf_{22}$
$t_3$	$tf_{31}$	$tf_{32}$

$(tf_{ij} \geq 0)$

该权重认为，  
项频越高，包含的  
信息量越多。



## 2. 向量空间模型

✓ 逆（倒）文档频率（inverse document frequency, IDF）：

$$w_i = idf_i = \log \frac{N}{df_i}$$

整个语料集有  $N$  个文档。 $df_i$  是指包含特征项  $t_i$  的文档数目。

文档 特征项	$d_1$	$d_2$
$t_1$	$df_{11}$	$df_{12}$
$t_2$	$df_{21}$	$df_{22}$
$t_3$	$df_{31}$	$df_{32}$

$idf$  是反映特征项在整个语料中重要性的全局性统计特征。一个特征项在某个文档中出现的频率越高，其包含的有效信息越低。

## 2. 向量空间模型

✓ 特征频率-逆文档频率 (TF-IDF) :  $tf\_idf_i = tf_i \times idf_i$

TF-IDF反映的思想是：区别文本最有意义的特征项应该是那些在当前文本中出现频率足够高，而在文本集合的其他文本中出现频率足够小的特征项。

## 2. 向量空间模型

### ◆ 举例

有如下文本：

人工 智能 是 计算  
机 科学 的 一个 分  
支， 它 企图 生产  
出 一种 能 以 人类  
智能 相似 的 方式  
作出 反应 的 智能  
机器。

词频：

0	教育
3	智能
1	人类
0	体育
0	足球
0	运动会
0	AI
⋮	
1	科学
0	文本
1	人工
1	计算机

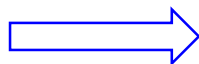
## 2. 向量空间模型

问题：文本越长项频越高。

词频：

有如下文本：

人工 智能 是 计算机 科学  
的 一个 分支 ， 它 企图  
生 产 出 一 种 能 以 人 类  
智 能 相 似 的 方 式 作 出  
反 应 的 智 能 机 器 。 人 工  
智 能 是 计 算 机 科 学 的  
一 个 分 支 ， 它 企 图 生  
产 出 一 种 能 以 人 类 智  
能 相 似 的 方 式 作 出 反  
应 的 智 能 机 器 。



0	教育
6	智能
2	人类
0	体育
0	足球
0	运动会
0	AI
:	
2	科学
0	文本
2	人工
2	计算机

## 2. 向量空间模型

### ◆ 文本长度规范化

对于文本:  $\mathbf{d} = (w_1, w_2, \dots, w_n)$ ,  $w_i$  为特征项的权重。

● 1-范数规范化:  $d_1 = \frac{d}{\|\mathbf{d}\|_1} = \frac{d}{\sum_i w_i}$

● 2-范数规范化:  $d_2 = \frac{d}{\|\mathbf{d}\|_2} = \frac{d}{\sqrt{\sum_i w_i^2}}$

● 最大词频规范化:  $d_{max} = \frac{d}{\|\mathbf{d}\|_\infty} = \frac{d}{\max_i \{w_i\}}$

## 2. 向量空间模型

### ◆ 向量空间模型的问题

- 采用什么特征项表示？
- 采用什么特征选择方法？
- 基于离散符号统计，表达不精细，无法反应类似特征之间的语义相似性。

# 本章内容

---

1. 问题提出

2. 向量空间模型

 3. 表示学习模型

4. 习题

# 3. 表示学习模型

## ◆两种代表性学习方法

- **文本概念表示模型**：以（概率）潜在语义分析和潜在狄利克雷分布（latent Dirichlet allocation, LDA）为代表的主题模型（topic model），旨在挖掘文本中隐含的主题或概念，文本被表示为主题分布的分布向量。
- **基于分布式表示的学习模型**：通过深度学习模型以最优化特定目标函数的方式在分布式向量空间中学习文本的低维实数向量表示。



# 3. 表示学习模型

## ◆不同粒度单位的表示学习

●词汇的表示学习

●短语的表示学习

●句子的表示学习

●动态的表示学习

word embedding

Word2Vector

➤ 基于语言模型的学习方法

➤ 直接学习方法

(见第6章)

❖ C&W Model

❖ CBOW and Skip-gram Model

❖ 负采样与噪声对比估计

❖ 字-词混合的表示学习

# 3. 表示学习模型

## ❖ C&W 模型

Collobert R and Weston J. 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning.  
In *Proceedings of ICML2008*, pages 160-177

## ✧ 基本思路

由上下文词预测当前词，使其概率最大(效果最优)。

# 3. 表示学习模型

◇ 举例说明

$$(w_i, Context) = w_{i-n}, \dots, w_{i-1}, \mathbf{w}_i, w_{i+1}, \dots, w_{i+n}$$

we have learned a **lot** from this lesson



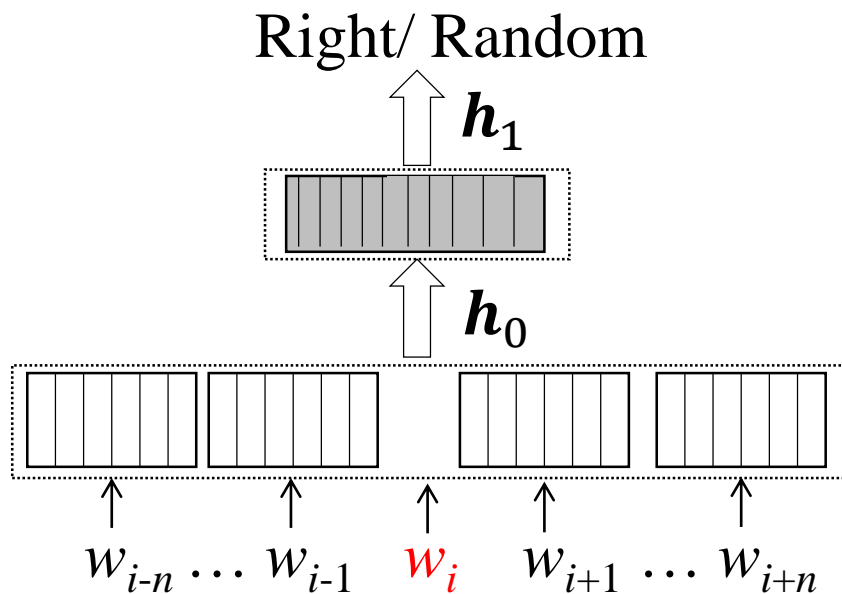
随机替换

$$(w'_i, Context) = w_{i-n}, \dots, w_{i-1}, \mathbf{w}'_i, w_{i+1}, \dots, w_{i+n}$$

we have learned a **book** from this lesson

$$score(w_i, Context) > score(w'_i, Context)$$

# 3. 表示学习模型



直接采用向量  
拼接的方法

$$h_0 = [e(w_{i-n}), \dots, e(w_{i-1}), e(w_i), e(w_{i+1}), \dots, e(w_{i+n})]$$

$$h_1 = f(W_0 h_0 + b_0)$$

得到 $n$ 元词组的得分:  $score(w_i, Context) = W_1 h_1 + b_1$


### 3. 表示学习模型

在向量优化过程中，C&W模型希望每个正样本的打分比对应的负样本的打分高1分，即：

$$score(w_i, Context) > score(w'_i, Context) + 1$$

对于整个训练语料，C&W模型需要遍历语料中的每个n元组，并最小化如下目标函数：

$$loss = \sum_{(w_i, C) \in D} \sum_{w' \in V'} \max(0, \underbrace{1 + score(w'_i, Context) - score(w_i, Context)}_{\text{red line}})$$



$$0 > score(w'_i, Context) + 1 - score(w_i, Context)$$

✧问题：上下文中的词顺序对预测结果有直接的影响。

# 3. 表示学习模型

## ❖ CBOW 模型

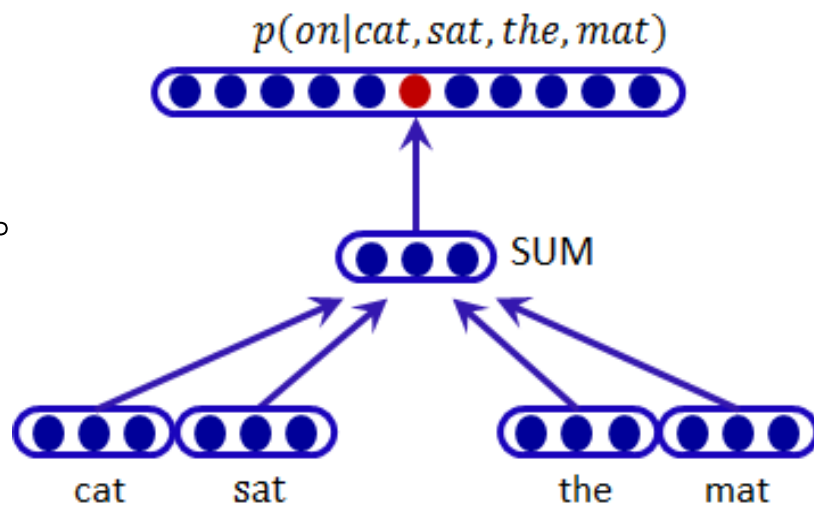
Tomas Mikolov, K. Chen et al . 2013. Efficient estimation of word representation in vector space. arXiv preprint arXiv: 1301.3781, 2013

### ✧ 基本思路

输入上下文词，预测中心目标词。

### ✧ 与C&W模型的区别

- 输入不再是上下文词对应词向量的拼接，直接采用所有词向量的平均；
- 省略隐层层，输入层直接与输出层链接，采用 Logistic 回归的形式计算中心目标词的概率。



# 3. 表示学习模型

给定训练语料中任意一个 $n$ 元组 ( $n = 2C+1$ ) :

$$(w_i, C) = w_{i-C} \dots w_{i-1} w_i w_{i+1} \dots w_{i+C},$$

将  $WC = w_{i-C} \dots w_{i-1} w_{i+1} \dots w_{i+C}$  作为输入, 计算上下文词的平均向量:

$$\mathbf{h} = \frac{1}{2n} \sum_{i-n \leq k \leq i+n, k \neq i} e(w_k)$$

$\mathbf{h}$  直接作为上下文的语义表示, 预测中心目标词  $w_i$  的概率:

$$p(w_i | WC) = \frac{\exp[\mathbf{h} \cdot e(w_i)]}{\sum_{k=1}^{|V|} \exp[\mathbf{h} \cdot e(w_k)]} \quad (\text{Softmax})$$

# 3. 表示学习模型

在CBOW模型中，词向量是唯一的神经网络参数。对于整个训练语料，模型优化词向量矩阵 $L_T$ ，以最大化所有词的对数似然：

$$\tilde{L}_T = \arg \max_{L_T} \sum_{w_i \in V} \log p(w_i | WC)$$



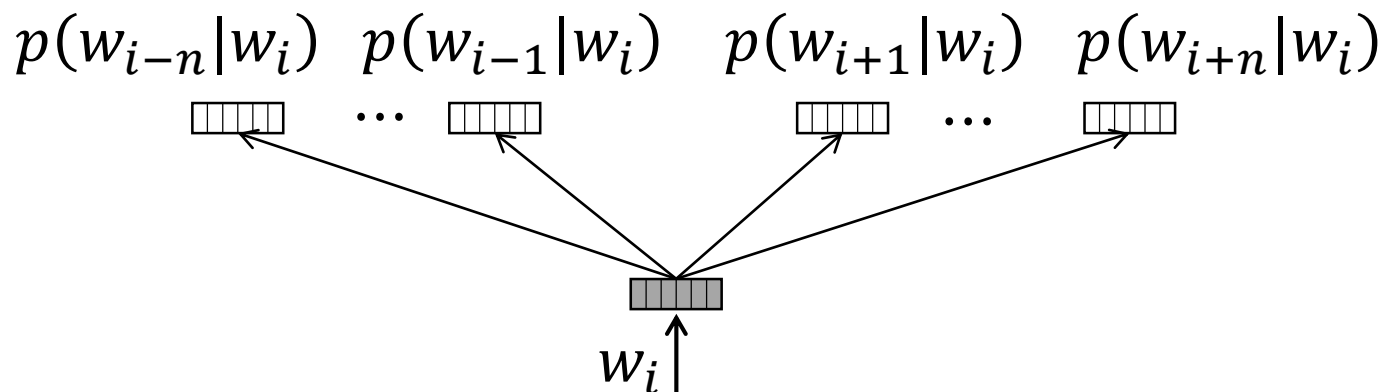
# 3. 表示学习模型

## ❖ Skip-gram 模型

Tomas Mikolov, K. Chen et al . 2013. Efficient estimation of word representation in vector space. arXiv preprint arXiv: 1301.3781, 2013

### ✧ 基本思路

用中心词预测所有的上下文词。



### 3. 表示学习模型

给定训练语料中任意一个 $n$ 元组 ( $n = 2C+1$ ) :

$$(w_i, C) = w_{i-C} \dots w_{i-1} w_i w_{i+1} \dots w_{i+C},$$

Skip-gram 模型直接利用中心词 $w_i$ 的词向量 $e(w_i)$ 预测上下文  
 $WC = w_{i-C} \dots w_{i-1} w_{i+1} \dots w_{i+C}$  作中每个词 $w_c$ 的概率:

$$p(w_c | w_i) = \frac{\exp[e(w_i) \cdot e(w_c)]}{\sum_{k=1}^{|V|} \exp[e(w_i) \cdot e(w_k)]}$$

Skip-gram 模型的目标优化函数与CBOW模型的目标函数类似，都是对于整个训练语料优化词向量矩阵 $L_T$ ，以最大化所有上下文词的对数似然：

$$\tilde{L}_T = \arg \max_{L_T} \sum_{w_i \in V} \sum_{w_c \in WC} \log p(w_c | w_i)$$

# 3. 表示学习模型

## ● 短语的表示学习

假设短语由 $i$ 个词语构成:  $ph_i = w_1 w_2 \dots w_i$

①视短语为词袋, 其表示为词向量的平均:

$$e(ph_i) = \sum_{k=1}^i e(w_k)$$

或者对词向量的每一维取最大:

$$e(ph_i) = \max_{k=1} (e(w_1), e(w_2), \dots, e(w_i))$$

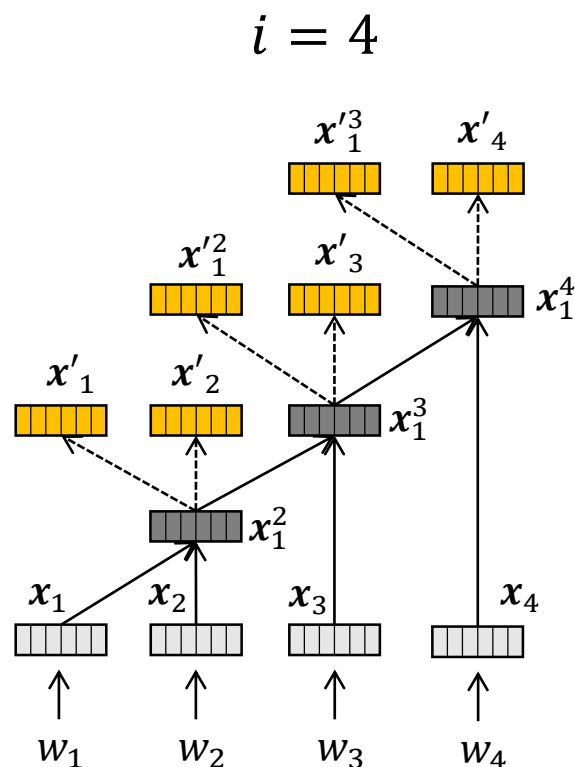
②视短语为词袋, 其表示为词向量的加权平均:

$$e(ph_i) = \sum_{k=1}^i v_k \cdot e(w_k)$$

# 3. 表示学习模型

## ③基于递归自编码器的表示学习(recursive autoencoder, RAE)

假设短语由 $i$ 个词语构成:  $ph_i = w_1 w_2 \dots w_i$



$$x_1^2 = f(W^{(1)}[x_1 : x_2] + W^{(1)})$$

$$\text{从 } x_1^2 \text{ 重构: } [x'_1 : x'_2] = f(W^{(2)}x_1^2 + W^{(2)})$$

$$E_{rec}([x_1 : x_2]) = \frac{1}{2} \|[x_1 : x_2] - [x'_1 : x'_2]\|^2$$

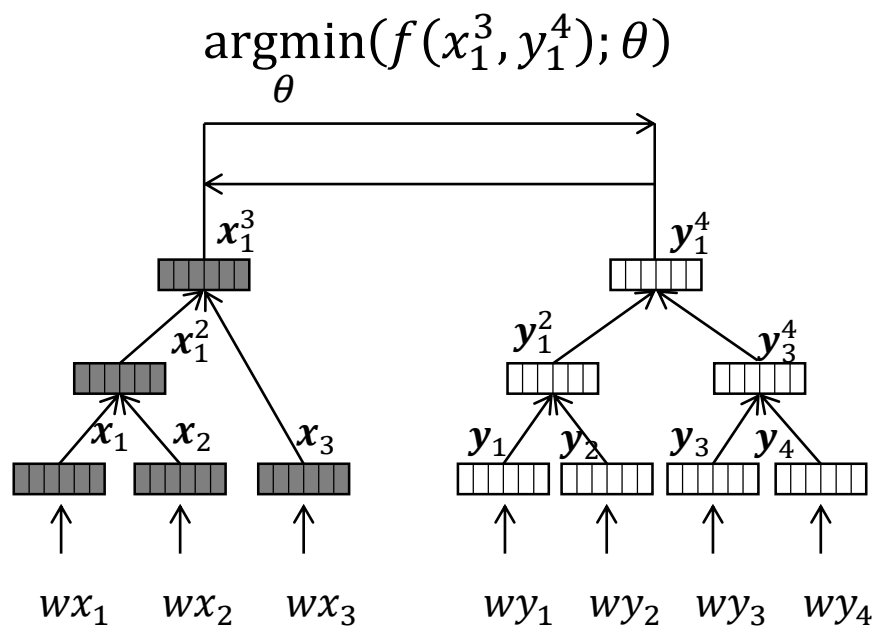
$$E_{\theta}(ph_i) = \operatorname{argmin}_{bt \in A(ph_i)} \sum_{nd \in bt} E_{rec}(nd)$$

二叉树集合

二叉树中的内部节点

# 3. 表示学习模型

➤ 验证：测试语义相近的短语在语义上向量空间能否聚在一起



假设用于训练的短语集合为  $S(ph)$ ，对于一个未见的短语  $ph^*$ ，利用短语向量之间的余弦距离度量任意两个短语之间的语义相似度，从  $S(ph)$  中搜索与  $ph^*$  相似的短语列表  $List(ph^*)$ ，检验  $List(ph^*)$  与  $ph^*$  是否真正语义相近。

# 3. 表示学习模型

➤ 例如

新输入短语	RAE
military force	core force main force labor force
at a meeting	to a meeting at a rate a meeting ,
do not agree	one can accept i can understand do not want
each people in this nation	each country regards each country has its each other , and

# 3. 表示学习模型

## ◆ 开源工具

- Google Word2Vec / word2phrase

<http://code.google.com/p/word2vec/>

# 本章内容

---

1. 问题提出
2. 向量空间模型
3. 表示学习模型

 4. 习题



## 4. 习题

1. 请分析说明短语表示学习是否可以采用词语表示学习方法（如CBOW和Skip-gram模型），为什么？
2. 利用北京大学标注的汉语分词语料，或者借助爬虫工具自己从互联网上收集足够多的英文语料，分别采用C&W、CBOW和Skip-gram模型得到汉语或英语的词向量，分析对比不同模型的结果差异。
3. 利用上面第2题的相同语料，利用 Google Word2Vec 开源工具获得汉语或英语的词向量表示，与第2题中自己实现的结果进行对比分析。

# 本章小结

## ◆ 向量空间模型

特征项；特征权重

## ◆ 分布式词表示学习

(1) C&W 模型

(2) CBOW 模型

(3) Skip-gram 模型

## ◆ 短语表示学习

谢谢!

*Thanks!*

