中国科学院大学
University of Chinese Academy of Sciences

自然语言处理

第4讲 词嵌入

王 石 资康莉 刘 瑜

2026年春季课程

https://ictkc.github.io/teaching/


国科大
UCAS
庚子
国科东路
UCAS East Road
南
SOUTH
北
NORTH
国科东路
UCAS East Road

第四讲 词嵌入

Transformer中我们忽略了啥?

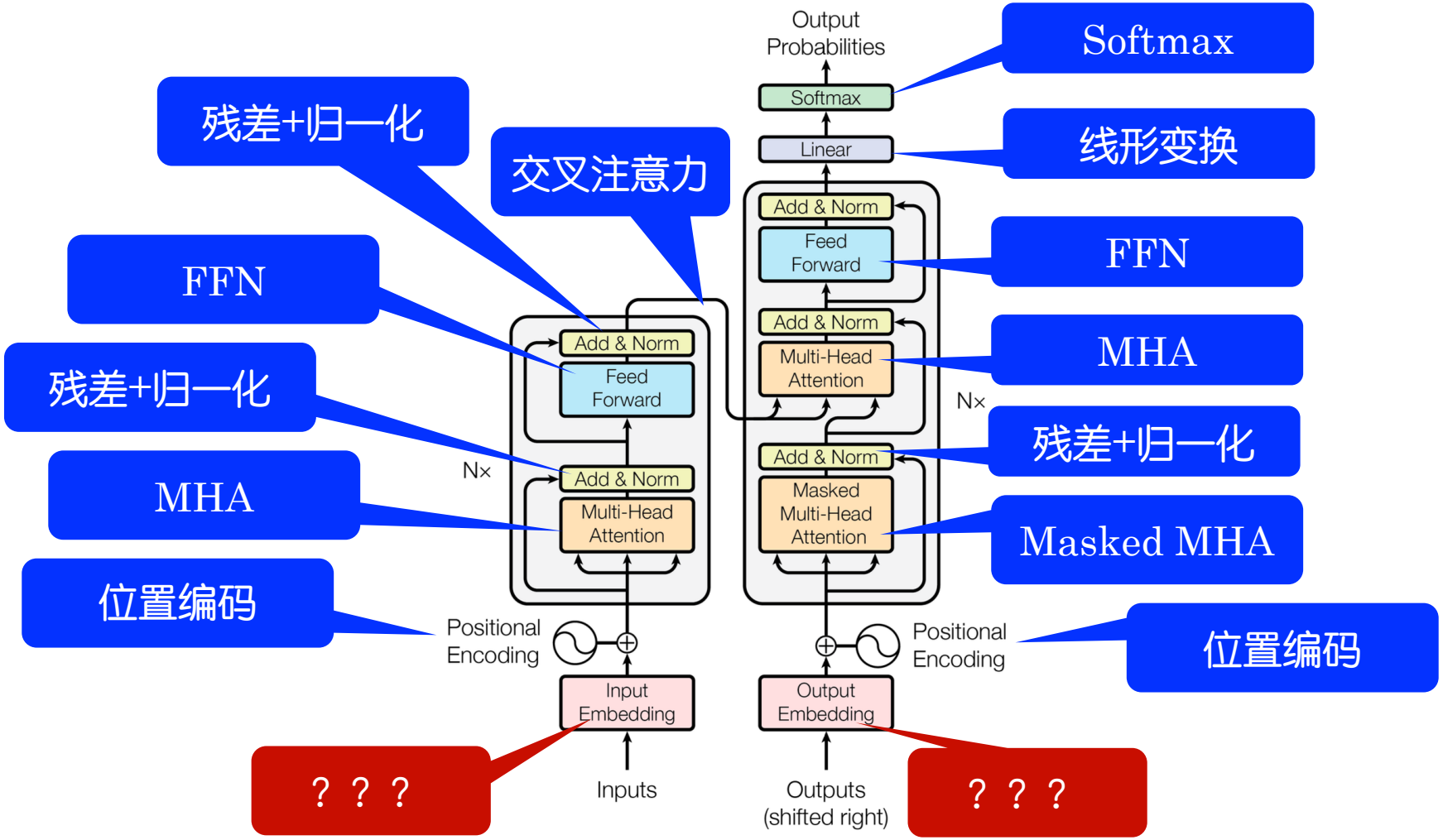


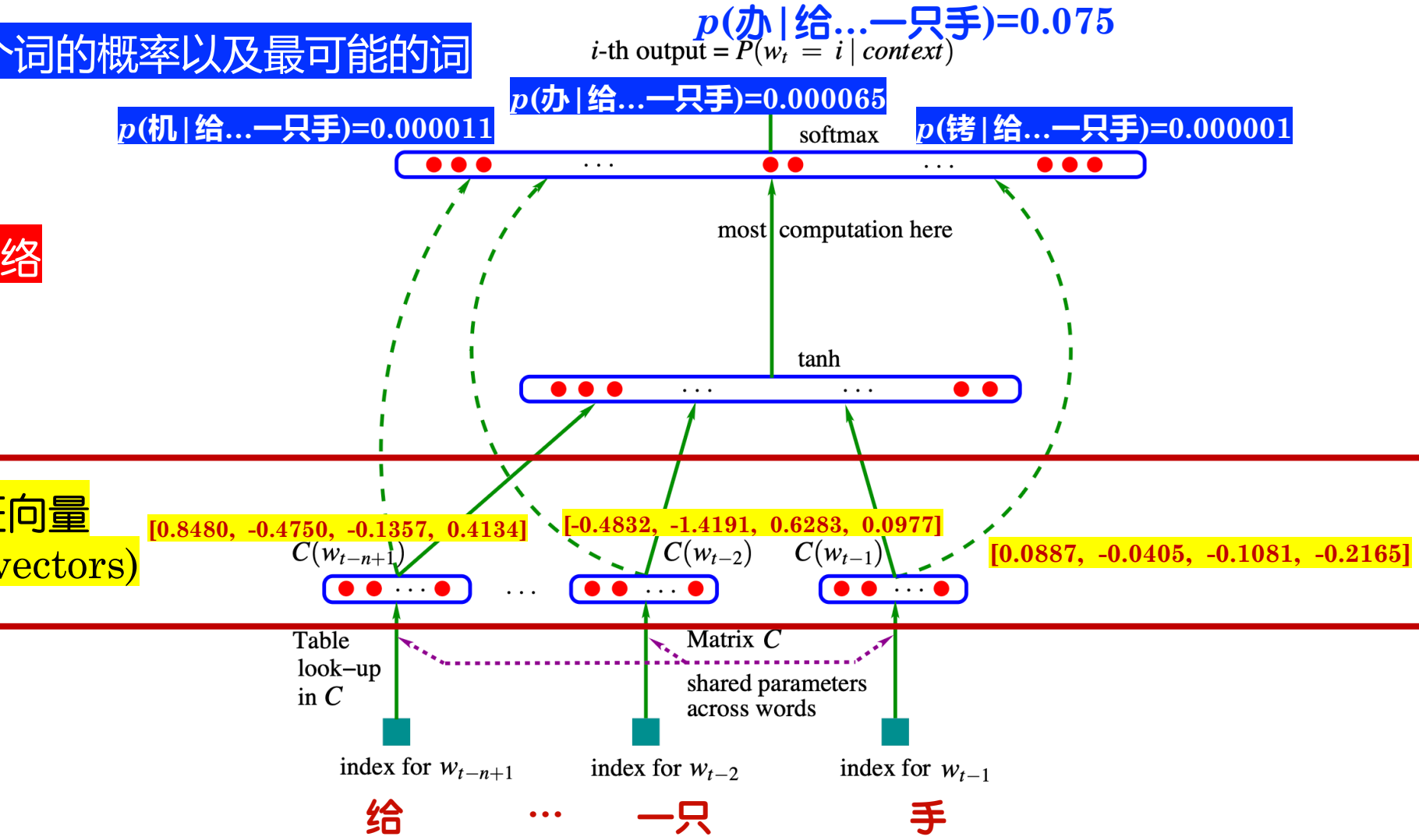
Figure 1: The Transformer - model architecture.

很早之前就见过

（3）输出每个词的概率以及最可能的词

（2）神经网络

（1）词分布式特征向量
(distributed feature vectors)

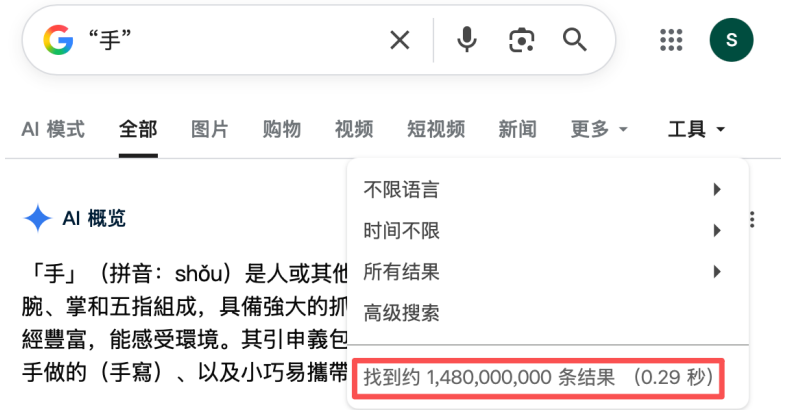


中国科学院大学
University of Chinese Academy of Sciences

目录

回顾：N-gram的弱泛化能力问题

- N-Gram Model：N越大，出现概率越低



解决方案：意思差不多得了

中文：同义词词林

Ba06B03= 家具 家电 农机具 食具 燃气具 灶具
Ba07A01= 礼品 礼物 礼 赠品 赠礼 人情 人事 仪 赐 贶 礼盒 礼金 红包 赠物
Ba07A02= 贺礼 贺仪
Ba07A03= 薄礼 小意思 千里鹅毛 谢礼
Ba07A04@ 厚礼
Ba07A05@ 见面礼
Ba07A06= 聘礼 彩礼 财礼
Ba07A07= 贡品 贡
Ba07A08= 祭礼 赙仪
Ba07A09# 寿礼 年礼 哈达
Ba07B01= 嫁妆 妆奁 妆 陪嫁 陪送
Ba07C01= 祭品 贡品 供 供品
Ba08A01= 宝物 宝贝 宝 珍 琛 珍品 珍宝 至宝 无价宝 瑰

https://github.com/BiLiangLtd/WordSimilarity/tree/master/data

中文：同义词词林

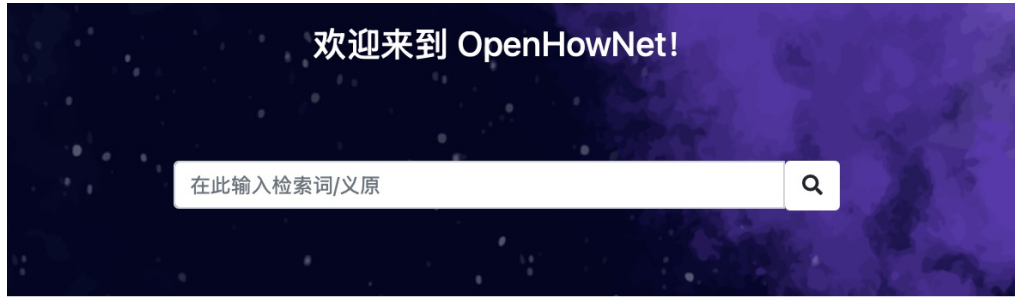
2 同义词词林介绍

《同义词词林》是梅家驹等人[5]于 1983 年编纂而成，这本词典中不仅包括了一个词语的同义词，也包含了一定数量的同类词，即广义的相关词。由于《同义词词林》著作时间较为久远，且之后没有更新，所以哈尔滨工业大学信息检索实验室利用众多词语相关资源，完成了一部具有汉语大词表的“哈工大信息检索研究室同义词词林扩展版”。《同义词词林扩展版》收录词语近 7万条，全部按意义进行编排，是一部同义类词典。哈工大信息检索研究室参照多部电子词典资源，并按照人民日报语料库中词语的出现频度，只保留频度不低于 3的 (小规模语料的统计结果）部分词语，剔除 14 706个罕用词和非常用词后，词表共包含 77 343条词语。

2.1 同义词词林分类方法

同义词词林按照树状的层次结构把所有收录的词条组织到一起，把词汇分成大、中、小 3类，大类有 12个，中类有 97个，小类有 1 400个。每个小类里都有很多的词，这些词又根据词义的远近和相关性分成了若干个词群 (段落)。每个段落中的词语又进一步分成了若干个行，同一行的词语要么词义相同 (有的词义十分接近)，要么词义有很强的相关性。例如，“大豆”、“毛豆”和“黄豆”在同一行；“西红柿”和“番茄”在同一行；“大家”、“大伙儿”、“大家伙儿”在同一行。另外，“将官”、“校官”、“尉官”在同一行，“雇农”、“贫农”、“下中农”、“中农”、“上中农”、“富农”在同一行，“外商”、“官商”、“坐商”、“私商”也在同一行，这些词不同义，但很相关。

中文：HowNet



特点

首次开源核心数据

点击了解知网

在线检索知网词条，展示义原结构

点击查看检索示例

提供丰富的调用接口方便用户使用

点击进入API项目页面

发展情况

237,974个中英文词条

35,202个概念

2,540个义原

30

构建时间近30年

https://openhownet.thunlp.org

OpenHowNet 基于义原的语言知识库

```
NO.=000000026417    # Concept ID
W_C=不惜     # Chinese word
G_C=verb    # POS tag of the Chinese word
S_C=PlusFeeling|正面情感  # Sentiment orientation
E_C=~牺牲业余时间，~付出全部精力，~出卖自己的灵魂    # Example s
W_E=do not hesitate to  # English word
G_E=verb    # POS tag of the English word
S_E=PlusFeeling|正面情感  # Sentiment orientation
E_E=                # Example sentences of the English
DEF={willing|愿意}   # Sememe-based definition
RMK=
```

英文：WordNet

WordNet的Synset

WordNet是一个英语词汇库。名词、动词、形容词、副词以及词组以同义group的形式放在一起，称为synsets，也就是一个概念(concept)。下面是一个称为synsets的例子：

```
mileage
fuel consumption rate
gasoline mileage
gas mileage
```

这就是一个synset，表示汽车的汽油里程数概念，同一个synset里的词可以看做是表达同一种意思，也就是同义词(synonym)。Synset也通过概念语义link连接在一起，使得整个Wordnet组成一个具有语义的网络。

英文：WordNet

WordNet Search - 3.1
- WordNet home page - Glossary - Help

https://wordnet.prince

Word to search for: dog Search WordNet

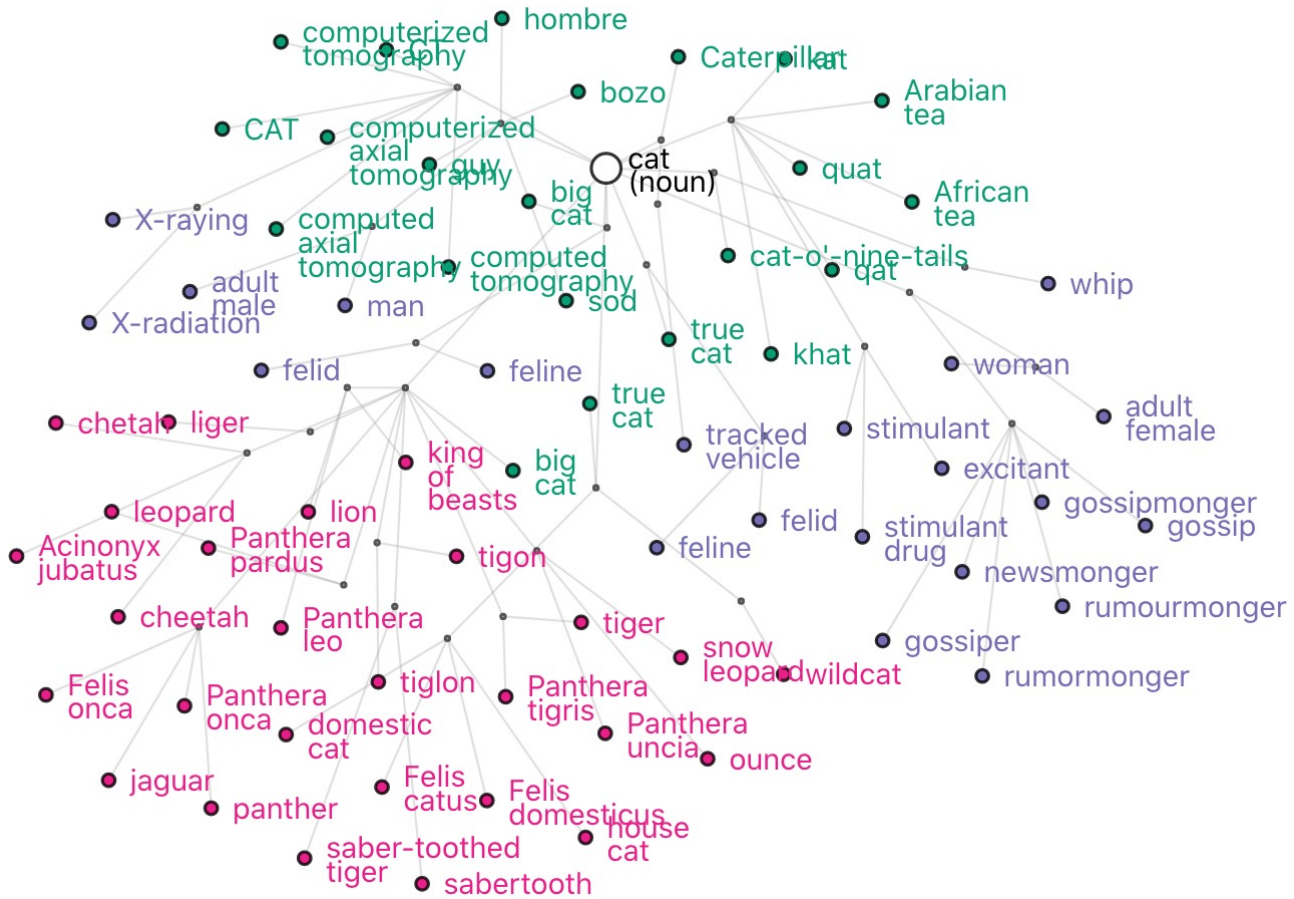
Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

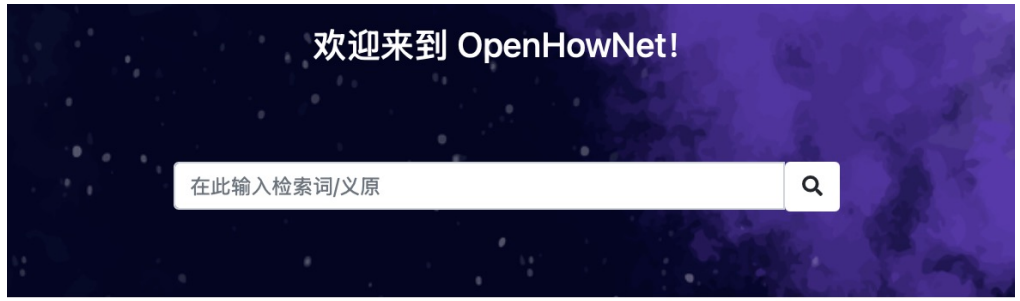
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **dog**, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) *"the dog barked all night"*
  - *direct hyponym* / *full hyponym*
  - *part meronym*
  - *member holonym*
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) canine, canid (any of various fissiped mammals with nonretractile claws and typically long muzzles)
      - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal) *"terrestrial carnivores have four or five clawed digits on each limb"*
        - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
          - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
            - S: (n) vertebrate, craniate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
              - S: (n) chordate (any animal of the phylum Chordata having a notochord or spinal column)
                - S: (n) animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)
                  - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
                    - S: (n) living thing, animate thing (a living (or once living) entity)
                      - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
                        - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
                          - S: (n) physical entity (an entity that has physical existence)
                            - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))



词典的缺点



特点

首次开源核心数据
点击了解知网

在线检索知网词条，展示义原结构
点击查看检索示例

提供丰富的调用接口方便用户使用
点击进入API项目页面

发展情况

237,974个中英文词条

35,202个概念

2,540个义原

30
构建时间近30年

❑ 难整理

家具 家电 农机具 食具 燃气具 灶具
礼品 礼物 礼 赠品 赠礼 人情 人事 仪 赐 贶 礼盒 礼金 红包 赠物
贺礼 贺仪
薄礼 小意思 千里鹅毛 谢礼
厚礼
见面礼
聘礼 彩礼 财礼
贡品 贡
祭礼 赙仪
寿礼 年礼 哈达
嫁妆 妆奁 妆 陪嫁 陪送
祭品 贡品 供 供品
宝物 宝贝 宝 珍 琛 珍品 珍宝 至宝 无价宝 瑰

这是我的一点**心意**

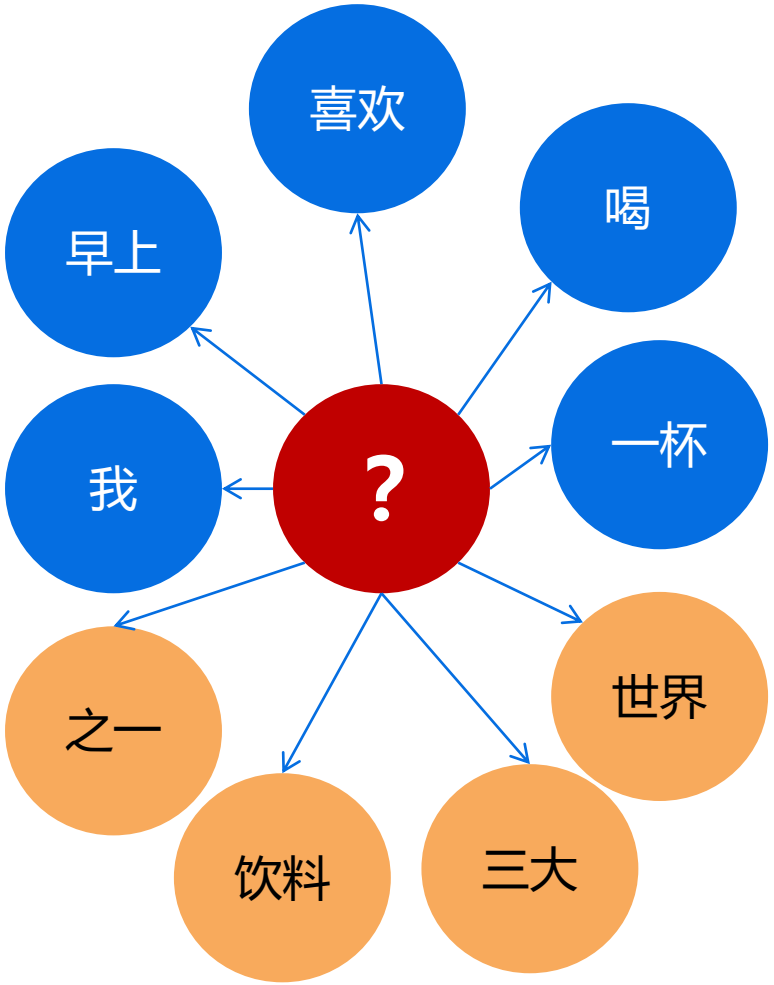
是大自然的**馈赠**

❑不全面

还有什么办法?

分布式假设（Distributional Hypothesis）

- Words that occur in the **same contexts** tend to have **similar meanings**[1]
- A word is **characterized by the company it keeps**[2]



[1] Harris, Z. (1954). Distributional structure.
[2] Firth, J.R. (1957). A synopsis of linguistic theory

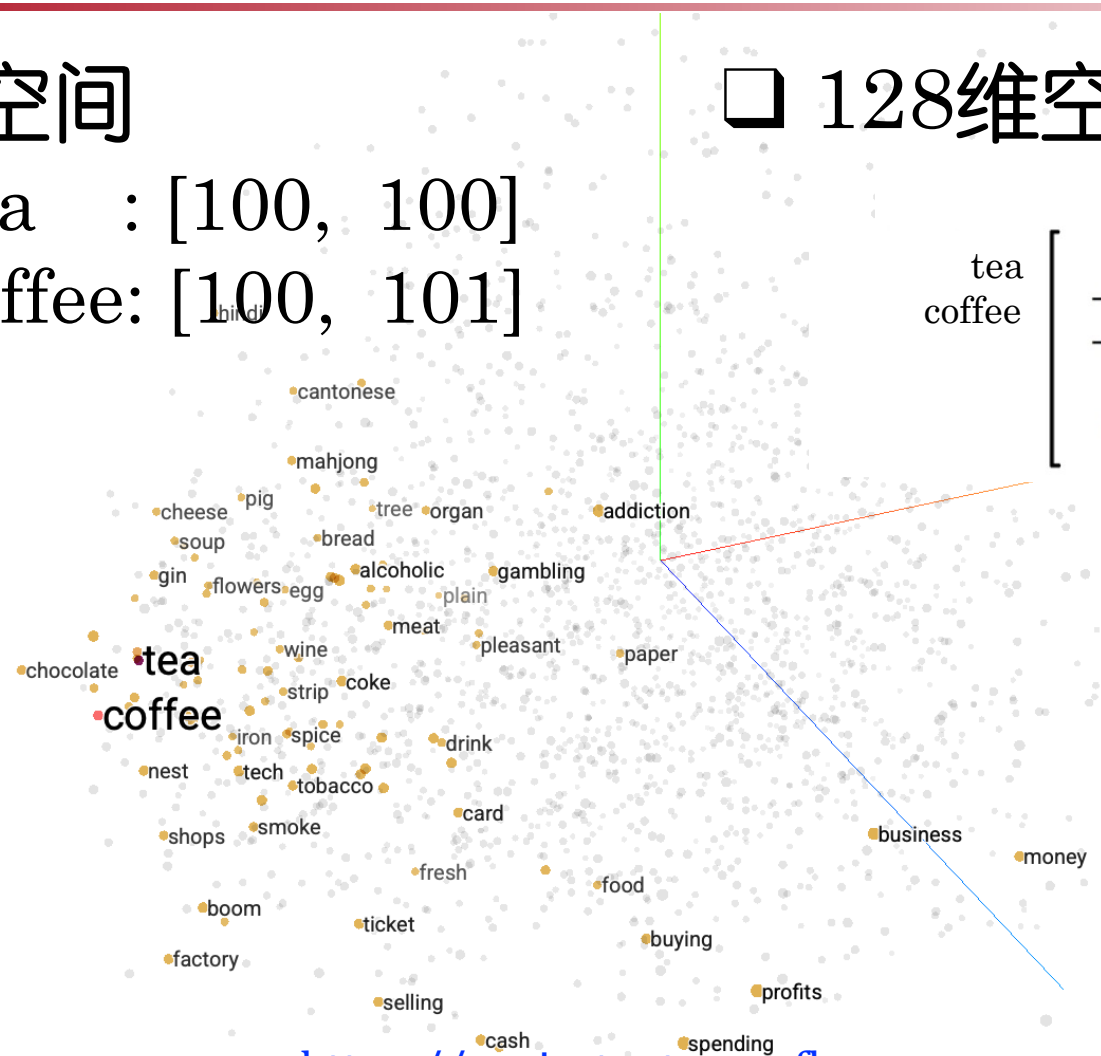
孟母三迁


孟母三迁
市
一迁墓地
二迁市集
三迁学宫
豆包AI生成


济宁市
1 邹城市北宿镇凫村
孟子故宅
凫村站
小元村
尚寨村
大元村
明鲁王墓
石柱厂村
大赵村
白马厂村
白马厂车站
京福线
路庄
小东章新村
后南宫村
西付村
10公里
后屯村
前南宫村
后八村
岗山铺
孟庄
2 邹城市凫山街道庙户营村
孟母三迁祠
前庙村
马庄
邹城站
邹城
邹城市
3公里
3 邹城市区南门外，孟庙一带
子思书院附近
孟庙孟府
护驾山公园

坐标化：词的分布式语义表示方法

- 2维空间
  - tea : [100, 100]
  - coffee: [100, 101]



- 128维空间

	第1维	第2维	第3维		第128维
tea	1.38192	1.82215	-1.25060	...	0.19055
coffee	-0.67227	-0.28646	0.47104	...	-0.01183
	-1.51263	-0.31815	-1.21959	...	-1.67615
	...	...	...	...	...
	0.30669	1.5996	0.21065	...	-1.85404

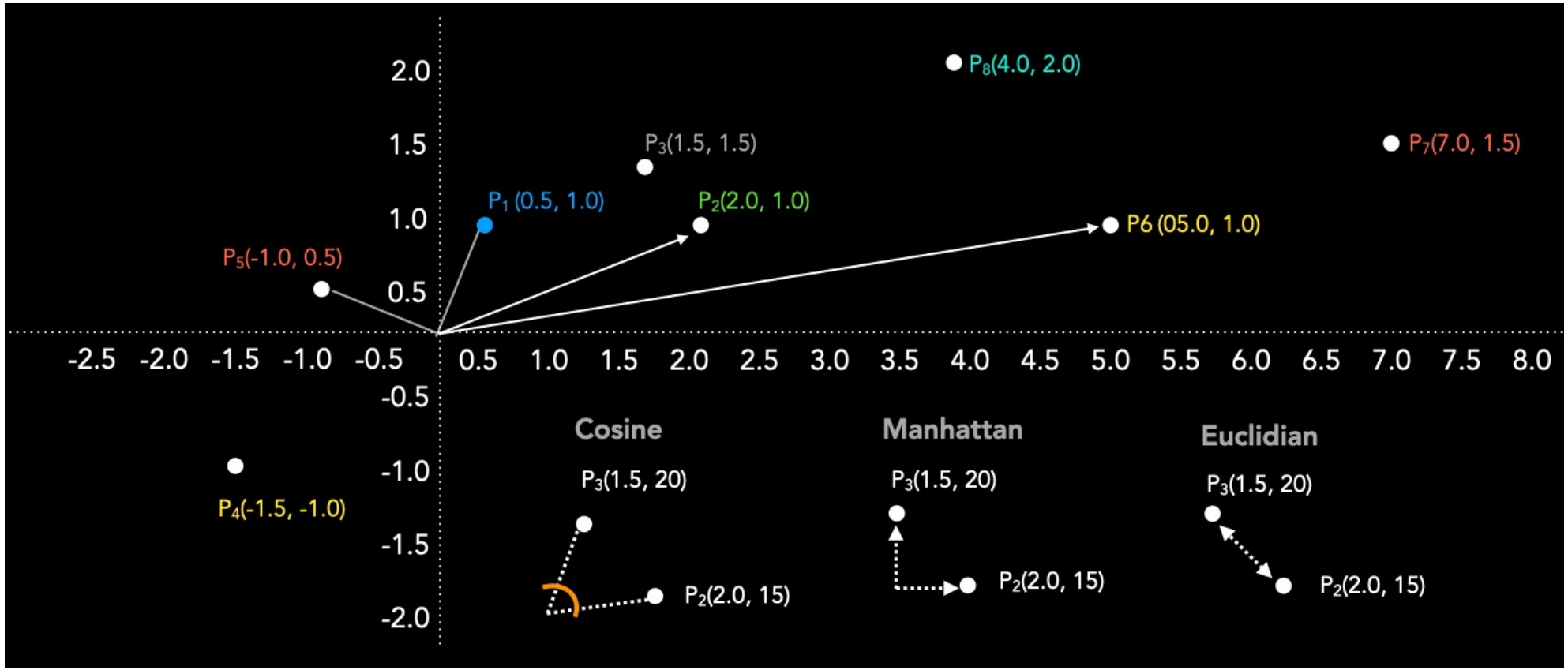
词向量

word vector

https://projector.tensorflow.org

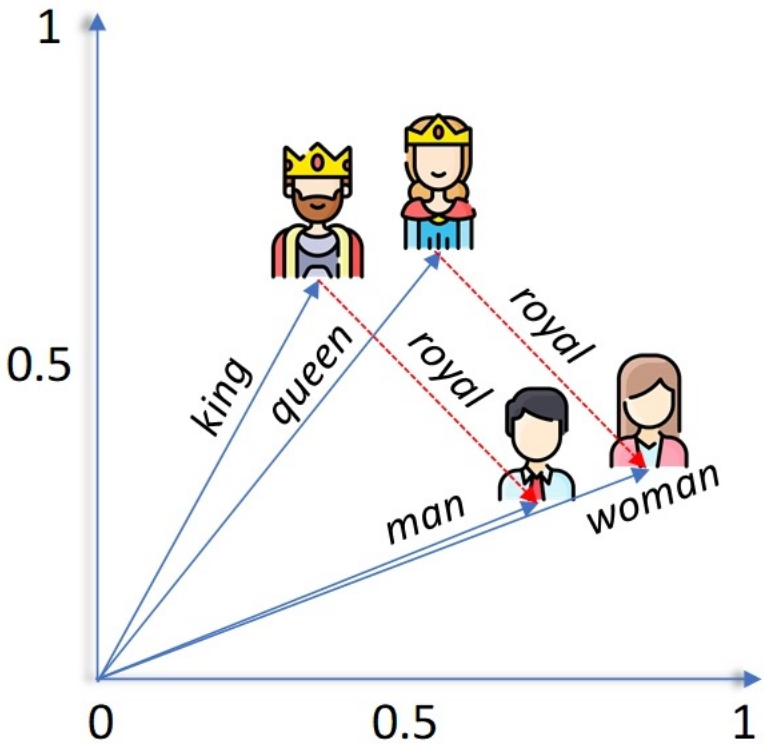
词向量的特征

- 用（余弦）空间距离表示语义距离

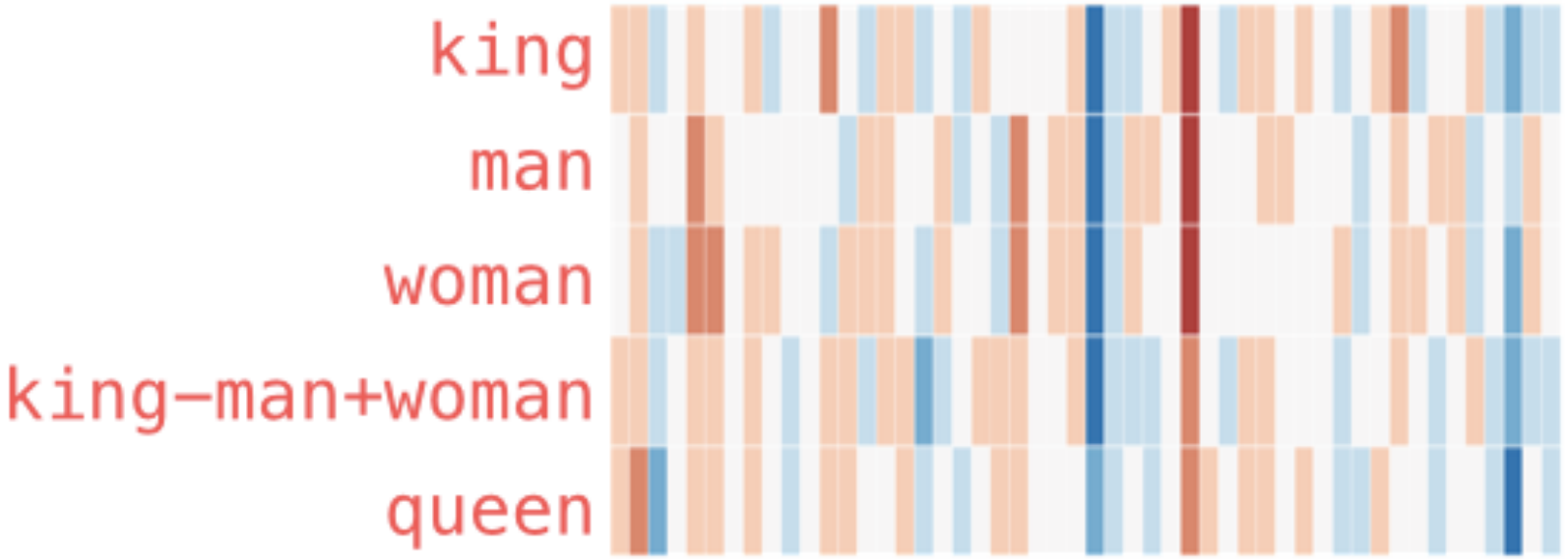


词向量的特征

- 可以通过向量运算“计算”语义

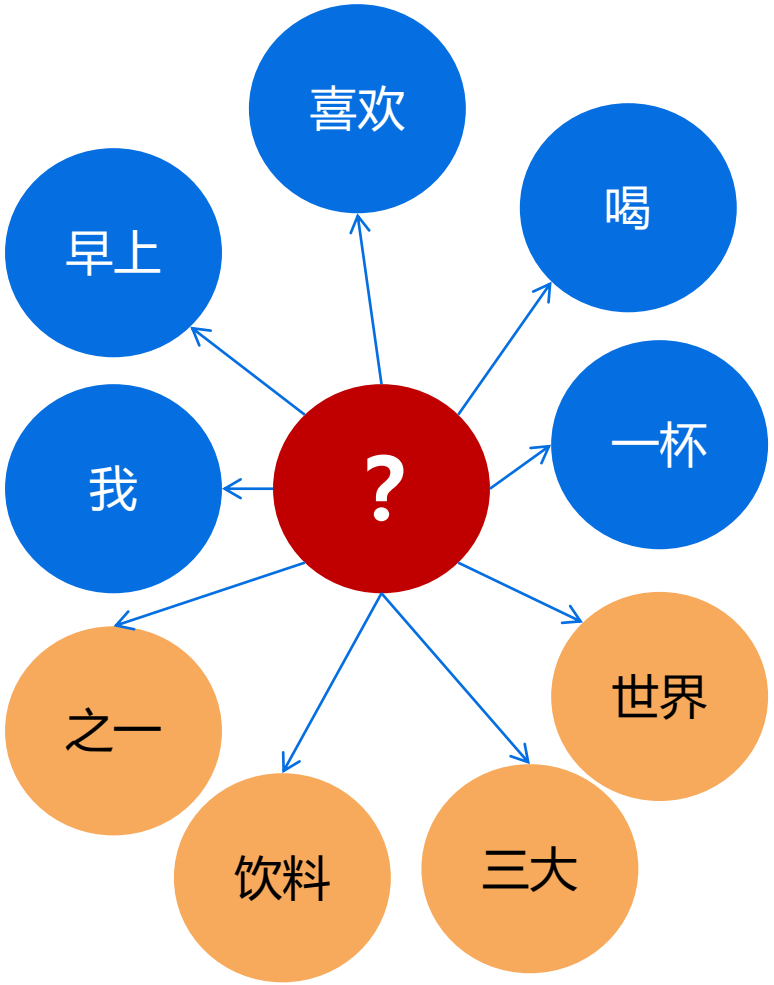


king − man + woman ~= queen



Embedding：得到词向量的过程

- Words that occur in the **same contexts** tend to have **similar meanings**[1]
- A word is **characterized by the company it keeps**[2]



中国科学院大学
University of Chinese Academy of Sciences

目录

怎么把一个词编码?

朱元璋 → **朱重八**

朱兴隆 → **朱重五**

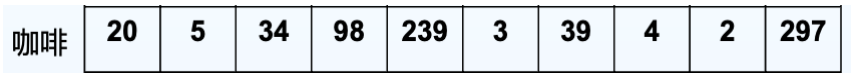
朱兴盛 → **朱重六**

朱兴祖 → **朱重七**

数数法：将邻居出现词的频率记下来

- 向量长度：词典大小
- 每一维：邻居出现频率

$N \approx 56,000$



《现代汉语常用词表》包含约5.6万个词语

	我	早上	喜欢	喝	一杯	咖啡	茶	世界	三大	饮料	…
我	22	23	1	2	123	20	18	23	5	23	
早上	23	15	3	23	1	5	5	4	3	0	
喜欢	1	3	4	87	43	34	35	2	0	12	
喝	2	23	87	2	126	98	99	0	2	123	
一杯	123	1	43	126	7	239	250	1	2	654	
咖啡	**20**	**5**	**34**	**98**	**239**	**3**	**39**	**4**	**2**	**297**	
茶	**18**	**5**	**35**	**99**	**250**	**39**	**5**	**3**	**3**	**302**	
世界	23	4	2	0	1	4	3	5	5	6	
三大	5	3	0	2	2	2	3	4	3	65	
饮料	23	0	12	123	654	297	302	6	65	0	

“咖啡”周围出现了20次“我”

“茶”周围出现了18次“我”

数数法：将邻居出现词的频率记下来

- **问题**：词表太大
  - 中文常用词约**5.6w**，矩阵规模25亿，int需**10G内存**
- **方案**：仅保留最常用词

	早上	咖啡	茶	世界	三大	饮料	…
我	23	20	18	23	5	23	
早上	15	5	5	4	3	0	
喜欢	3	34	35	2	0	12	
喝	23	98	99	0	2	123	
一杯	1	239	250	1	2	654	
咖啡	**5**	**3**	**39**	**4**	**2**	**297**	
茶	**5**	**39**	**5**	**3**	**3**	**302**	
世界	4	4	3	5	5	6	
三大	3	2	3	4	3	65	
饮料	0	297	302	6	65	0	

再看NNLM

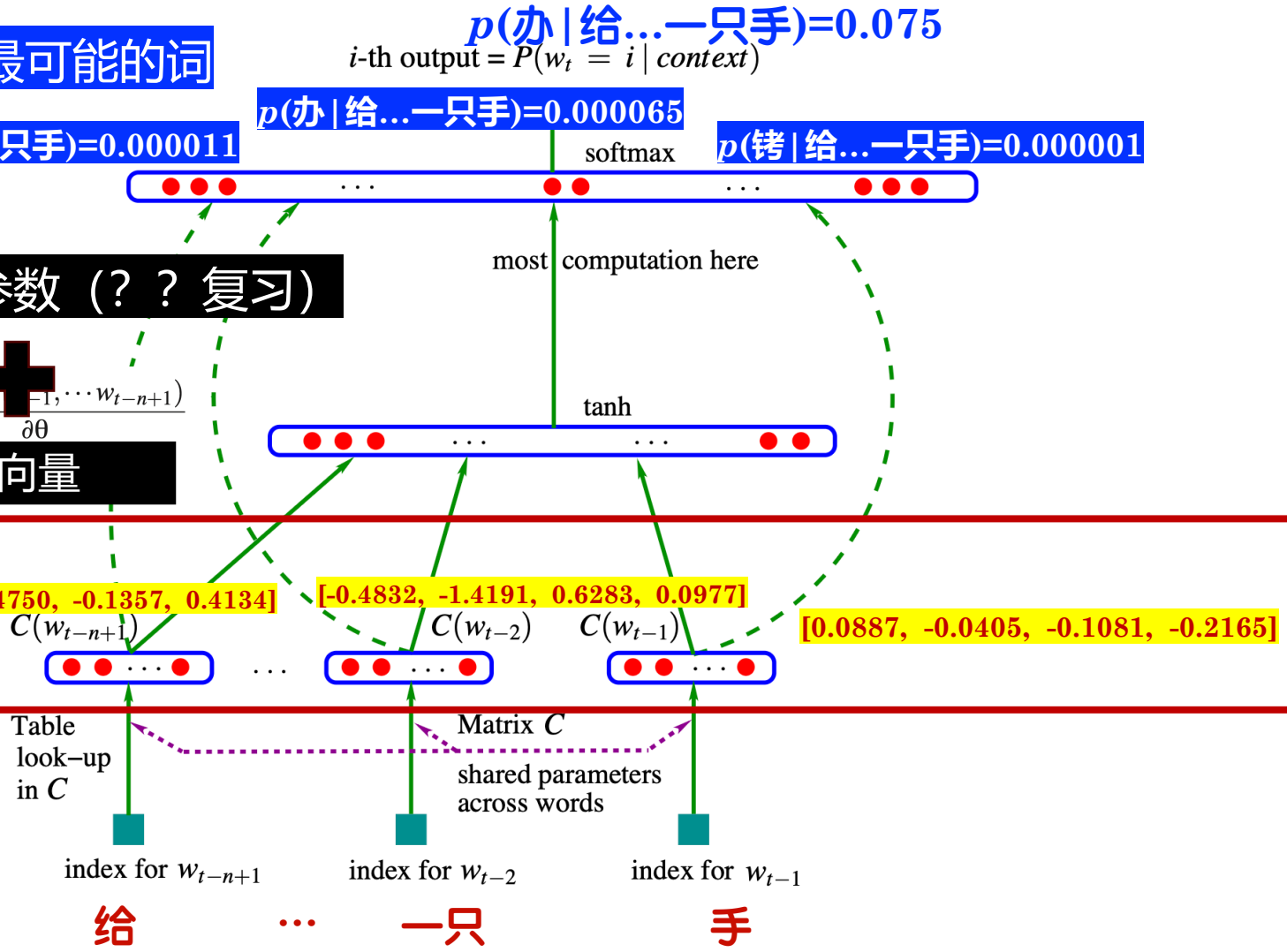
（3）输出每个词的概率以及最可能的词

（2）神经网络拟合：网络参数（？？复习）

$$L = \frac{1}{T}\sum_t \log f(w_t, w_{t-1}, \cdots, w_{t-n+1}; \theta) + R(\theta), \quad \theta \leftarrow \theta + \varepsilon \frac{\partial \log \hat{P}(w_t | w_{t-1}, \cdots w_{t-n+1})}{\partial \theta}$$

词向量

（1）词分布式特征向量
(distributed feature vectors)



词向量：NNLM的“副产品”

输入

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

预测

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

参数

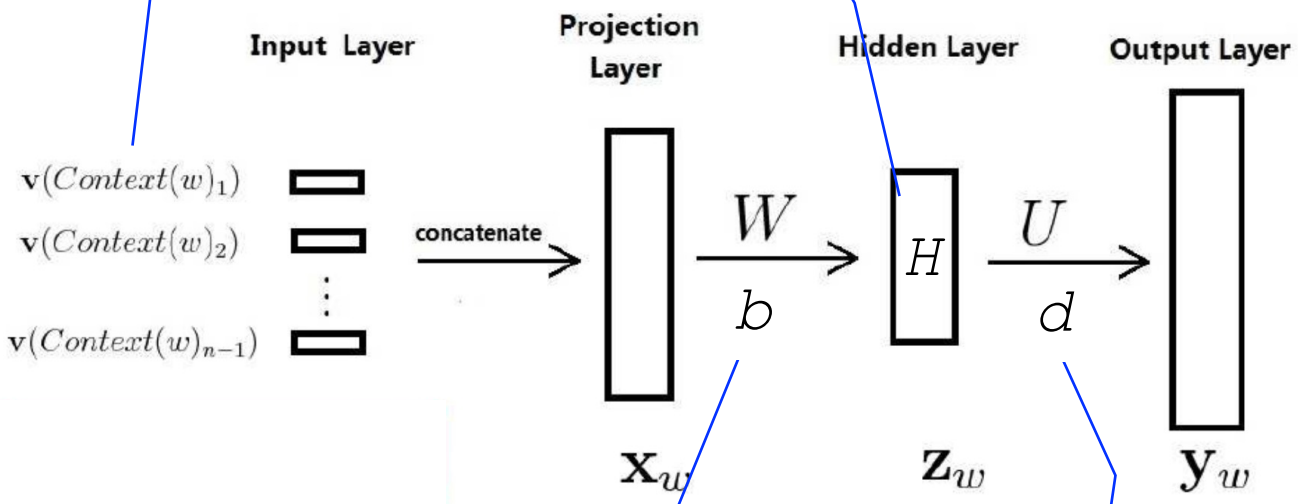
$$\theta = (b, d, W, U, H, C).$$

训练

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

词向量表示

H：隐层参数



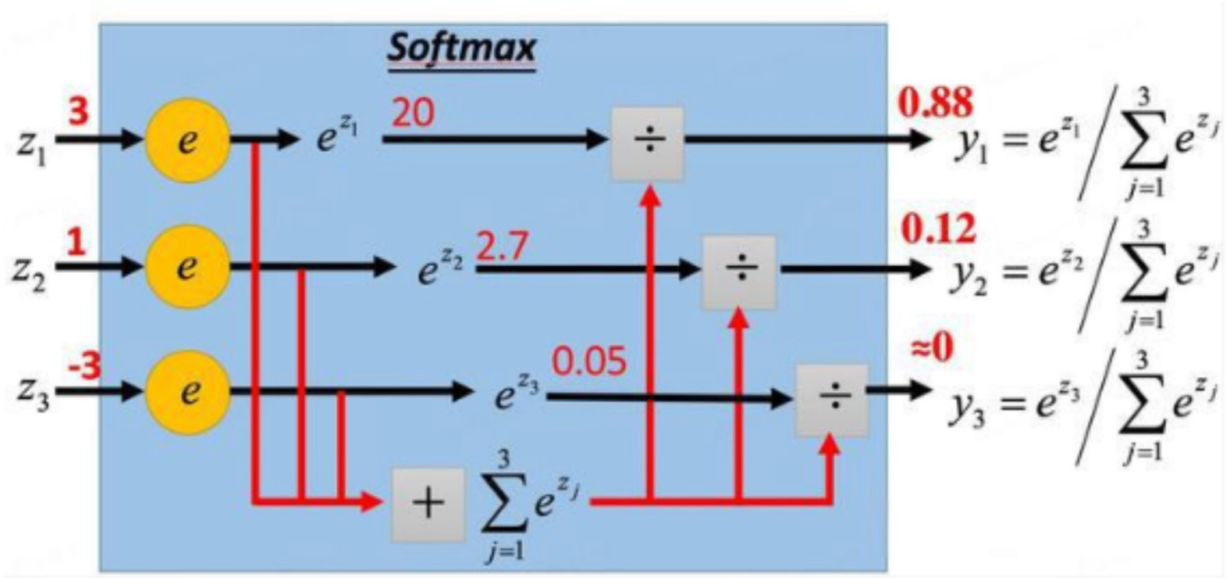
W：词表示层到隐层参数
b：输出偏置

U：隐层到输出层参数
d：隐层偏置

不足之处

1. 仅用了上文信息

2. 输出层Softmax计算量大



每次迭代需要的加、除次数：词表规模

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

从理论突破走向实际应用：Word2Vec (13')

❑ NNLM和Word2Vec的训练效率对比

Table 6: *Comparison of models trained using the DistBelief distributed framework. Note that training of NNLM with 1000-dimensional vectors would take too long to complete.*

Model	Vector Dimensionality	Training words	Accuracy [%]			Training time [days x CPU cores]
			Semantic	Syntactic	Total	
NNLM	100	6B	34.2	64.5	50.8	14 x 180
CBOW	1000	6B	57.3	68.9	63.7	2 x 140
Skip-gram	1000	6B	66.1	65.1	65.6	2.5 x 125

从理论突破走向实际应用：Word2Vec (13')

Distributed representations of words and phrases and their compositionality

T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean

Advances in neural information processing systems, 2013 • proceedings.neurips.cc

Abstract

The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several improvements that make the Skip-gram model more expressive and enable it to learn higher quality vectors more rapidly. We show that by subsampling frequent words we obtain significant speedup, and also learn higher quality representations as measured by our tasks. We also

展开

☆ 保存 引用 被引用次数: 49543 相关文章 所有 46 个版本 »

NIPS Test of Time Award

Tomas Mikolov

2023年12月14日 ·

Yesterday we received a Test of Time Award at NeurIPS for the word2vec paper from ten years ago. I'm really happy about it! I think it's the first "best paper" type of award I ever received. In fact, the original word2vec paper was rejected at the first ICLR conference in 2013 (despite the acceptance rate of around 70%), so it made me think how difficult it is for reviewers to predict future impact of research papers.

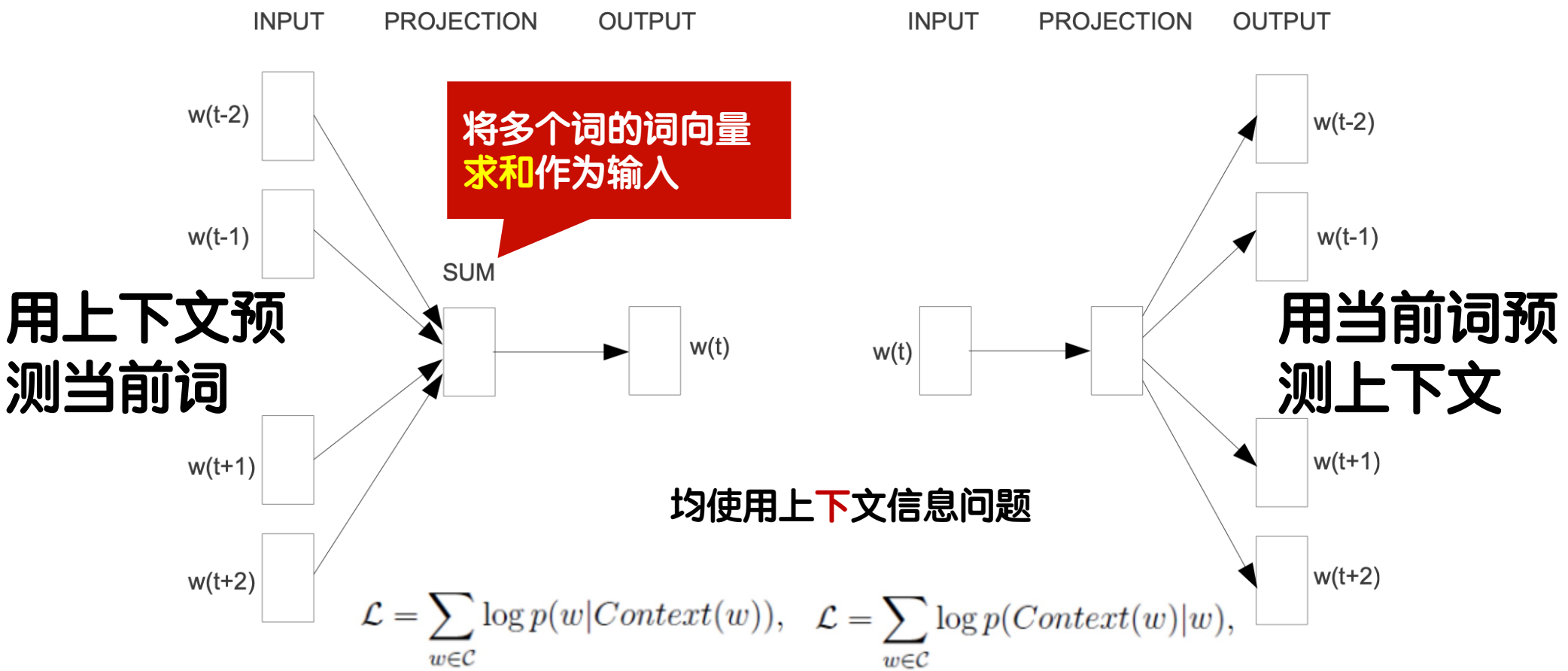
I've heard a lot of comments - both positive and negative - word2vec during those years, and did not really comment online about it. How I felt the research community is constantly flooded by propaganda certain researchers who are hacking this way the citation counts there, and I did not wa to share so

第一次投稿被ICLR2013拒

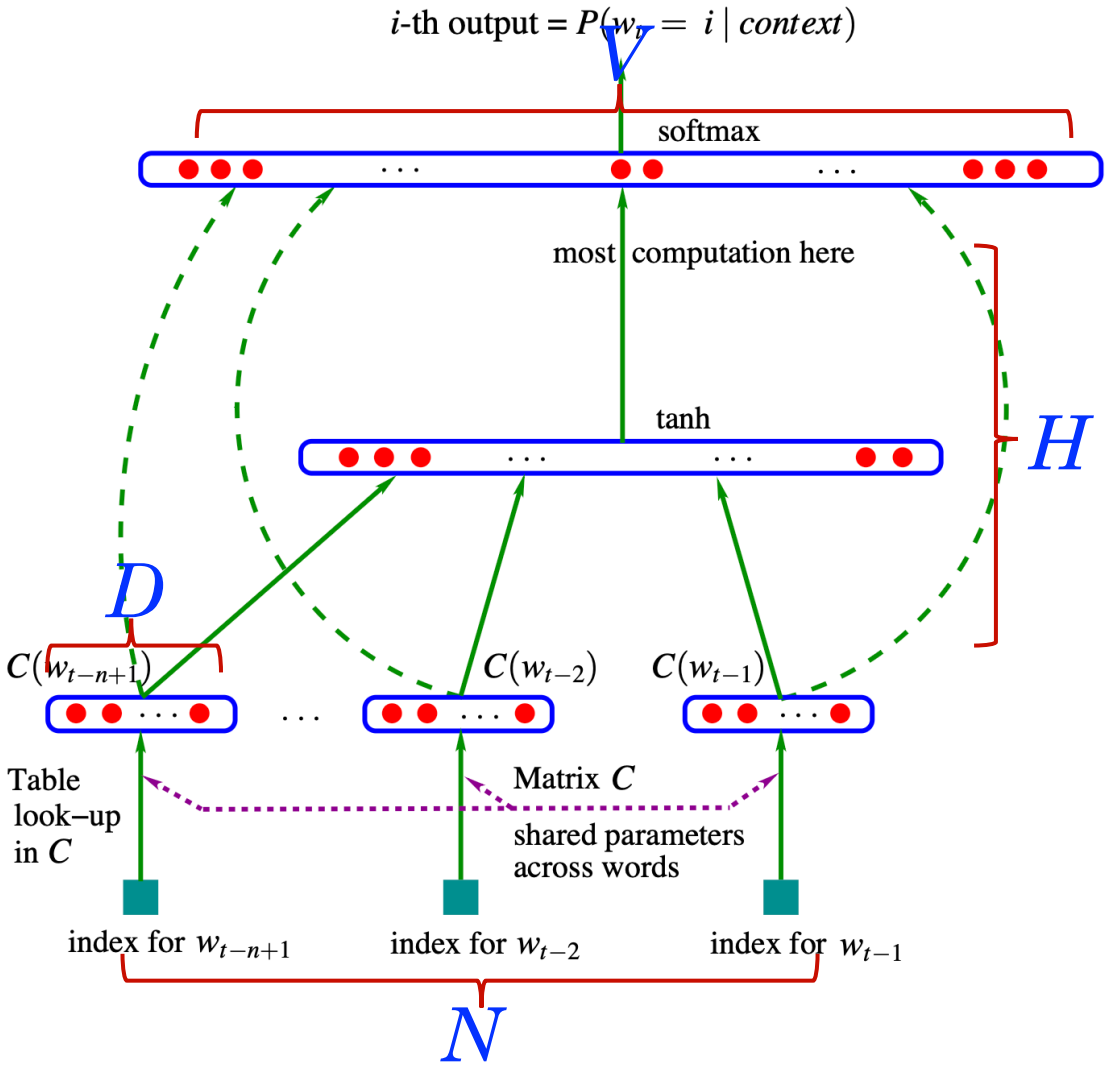
negatively surprised when they ended up publishing my idea under now famous name "sequence to sequence" where not only I was not mentioned as a co-author, but in fact my former friends forgot to mention me also in the long Acknowledgement section, where they thanked personally pretty much every single person in Google Brain except me. This was the time when money started flowing massively into AI and every idea was worth gold. It was sad to see the deep learning community quickly turn into some sort of Game of Thrones. Money and power certainly corrupts people...

感叹金钱和权利确实会孵化人心…

Word2Vec的两个模型



时间复杂度： NNLM



上下文词个数

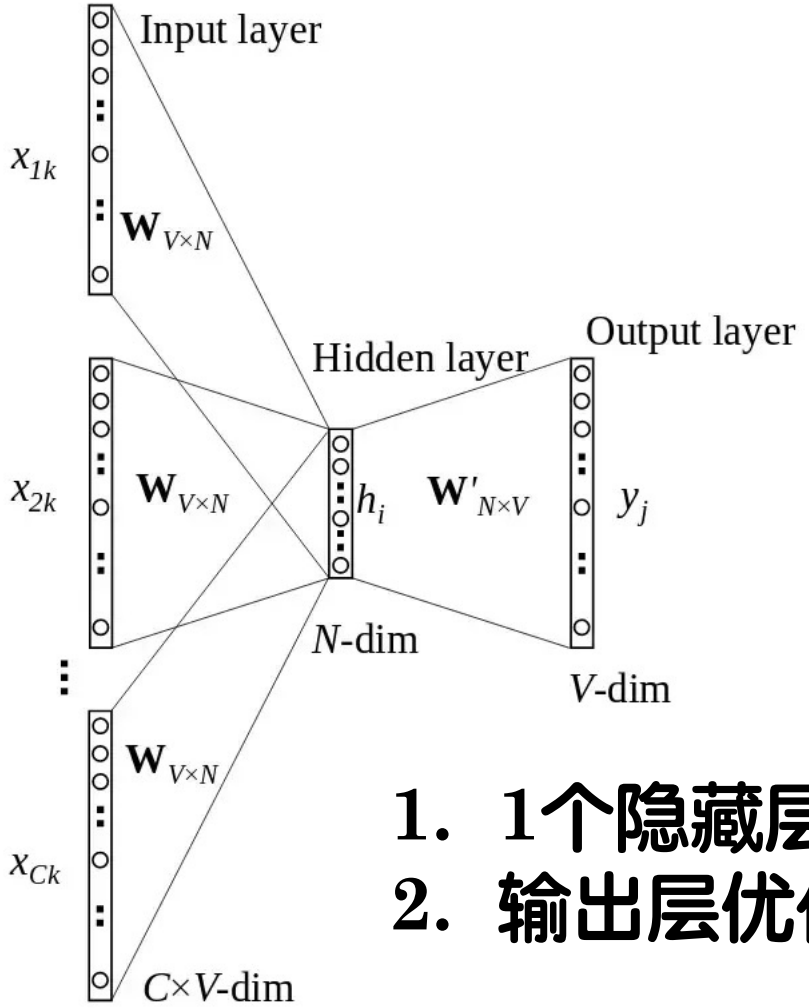
词表大小

$$Q = N \times D + N \times D \times H + H \times V,$$

词向量长度

隐层大小

时间复杂度： CBOW



1. 1个隐藏层
2. 输出层优化

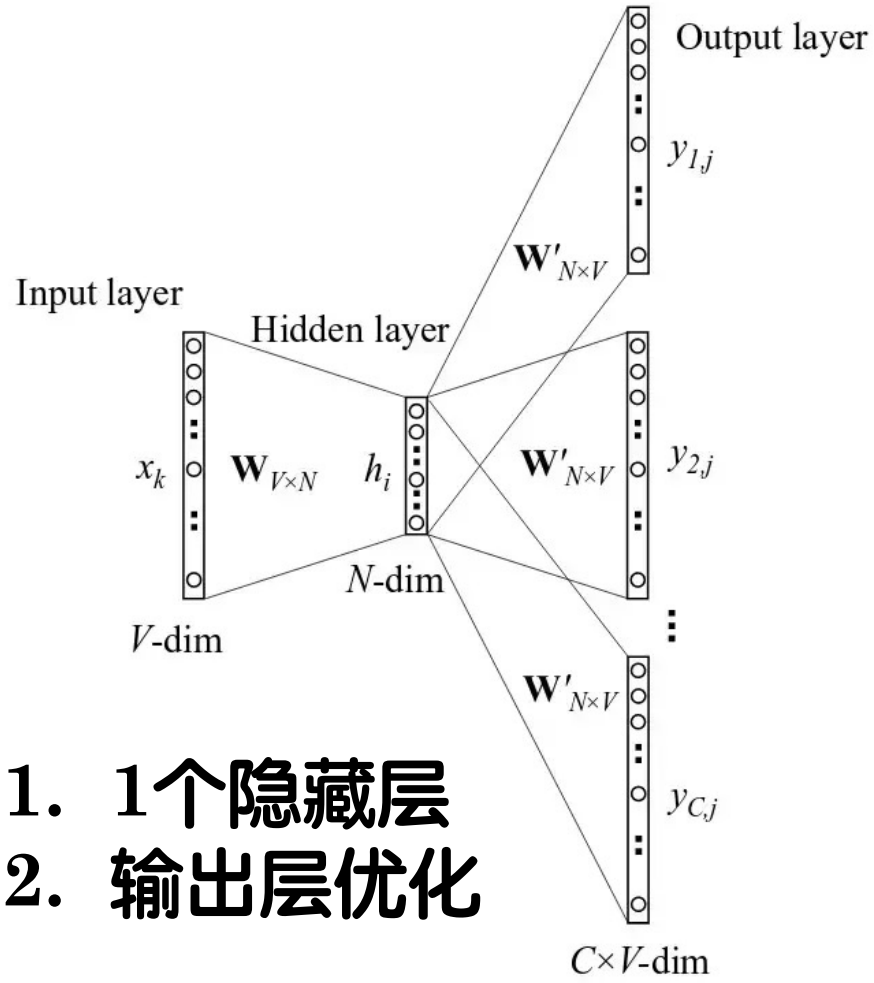
上下文词个数

词表大小

$$Q = N \times D + D \times log_2(V).$$

词向量长度

时间复杂度： Skip-Gram



1. 1个隐藏层
2. 输出层优化

上下文词个数

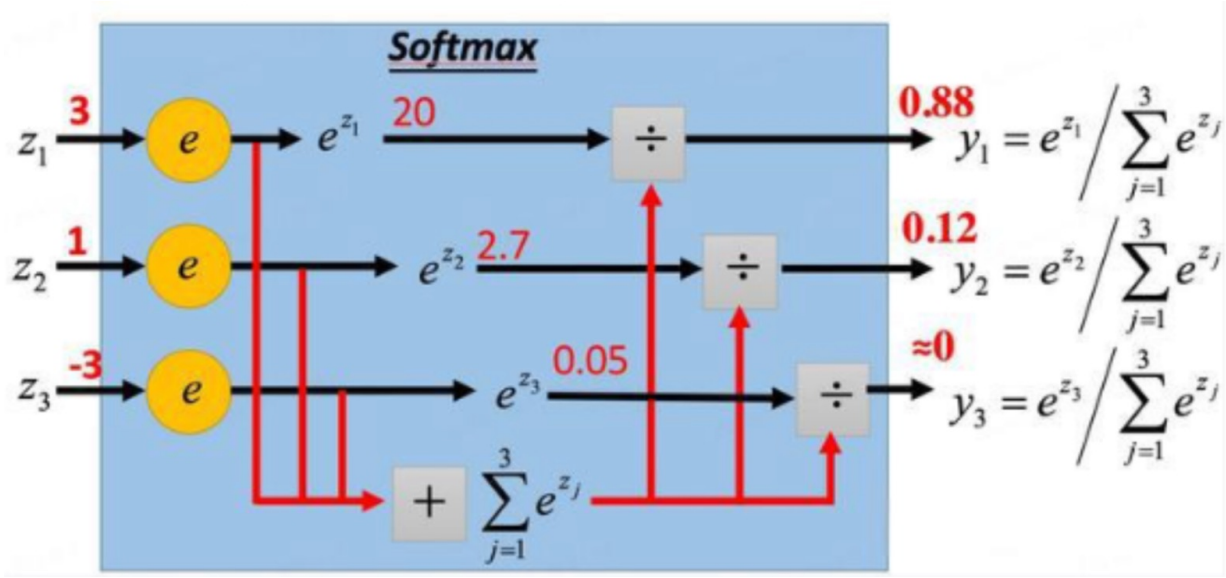
词表大小

$$Q = C \times (D + D \times log_2(V)),$$

词向量长度

NNLM：输出层复杂度为V

输出层Softmax计算量大



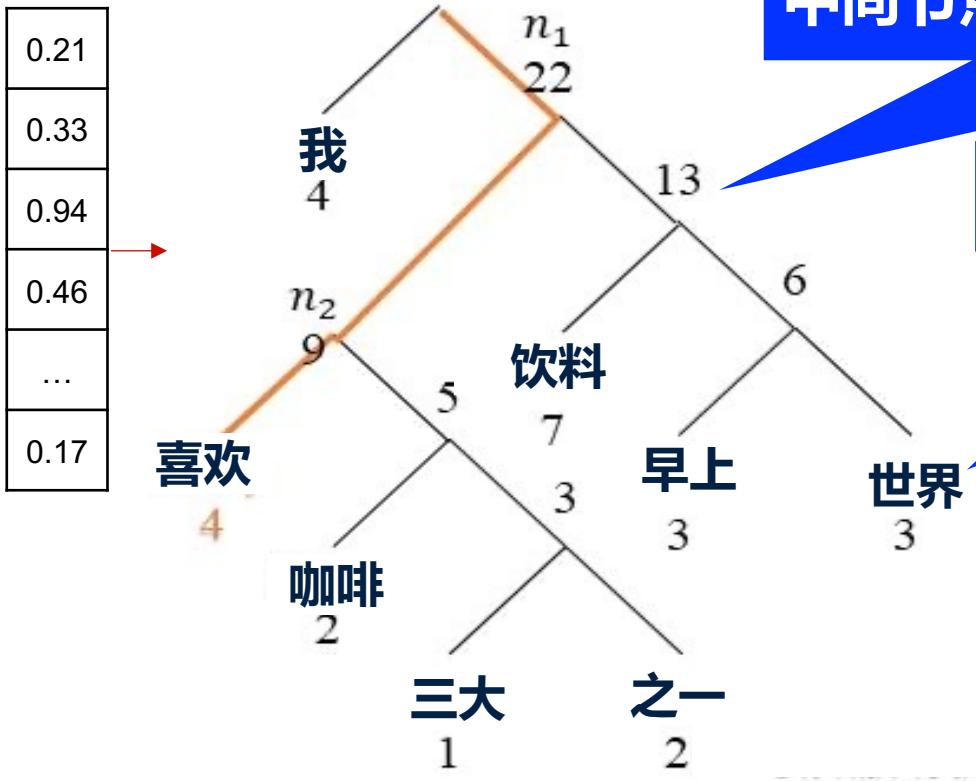
每次迭代需要的加、除次数：词表规模

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

Hierarchical Softmax: $log_2(V)$

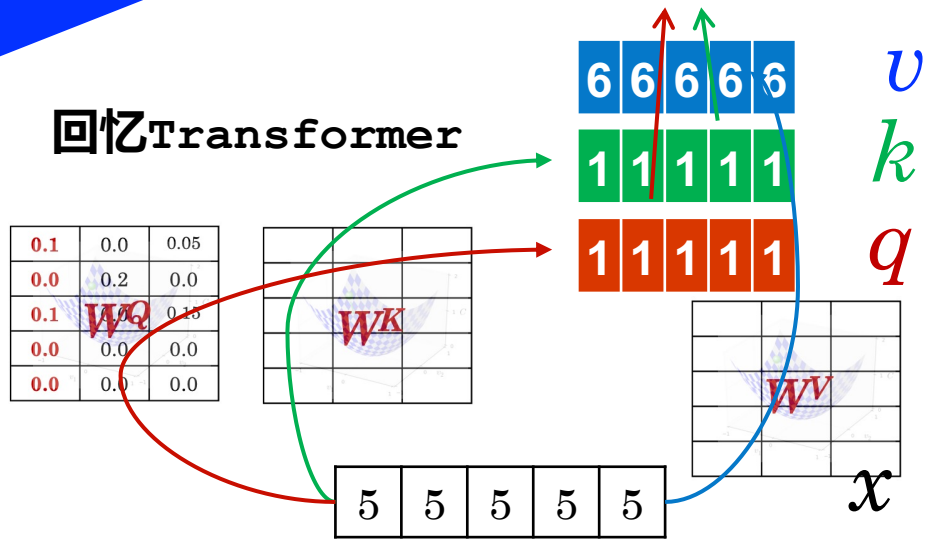
词频Huffman树

训练一组参数， 词向量与之相乘直接的到概率



中间节点为sigmoid二分类器，需训练参数

叶子节点为词，其概率为达到路径节点概率之积

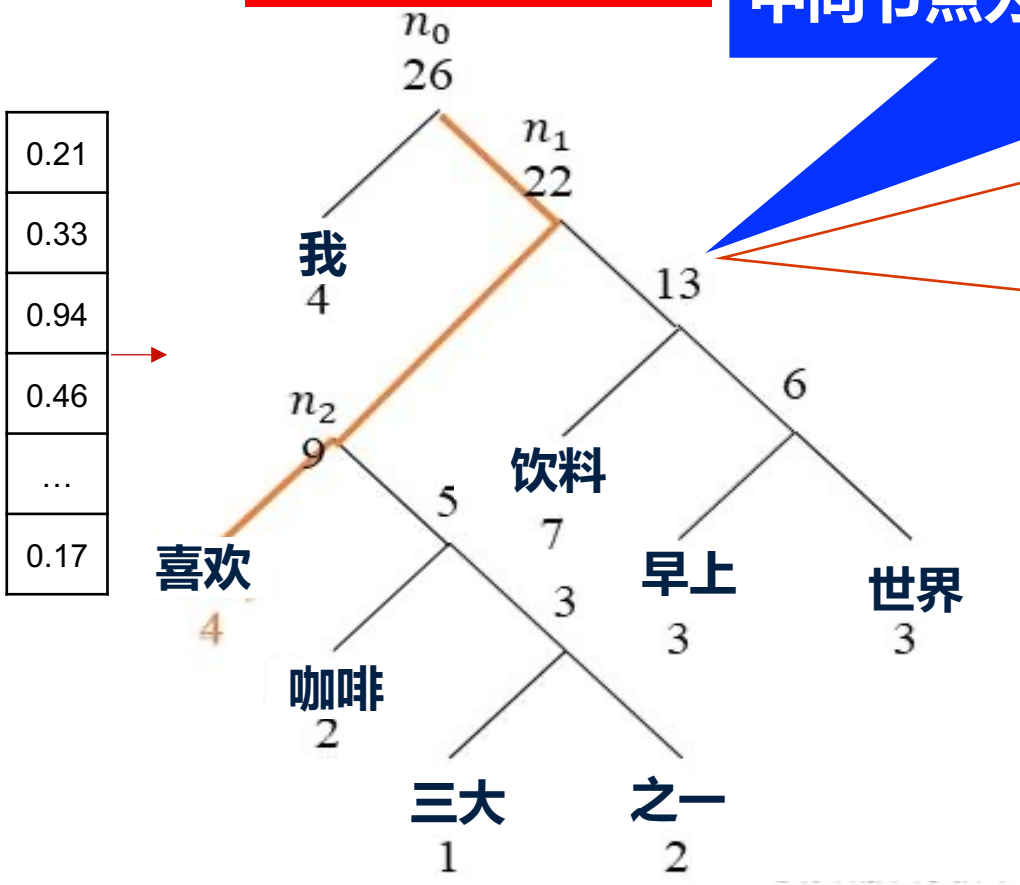


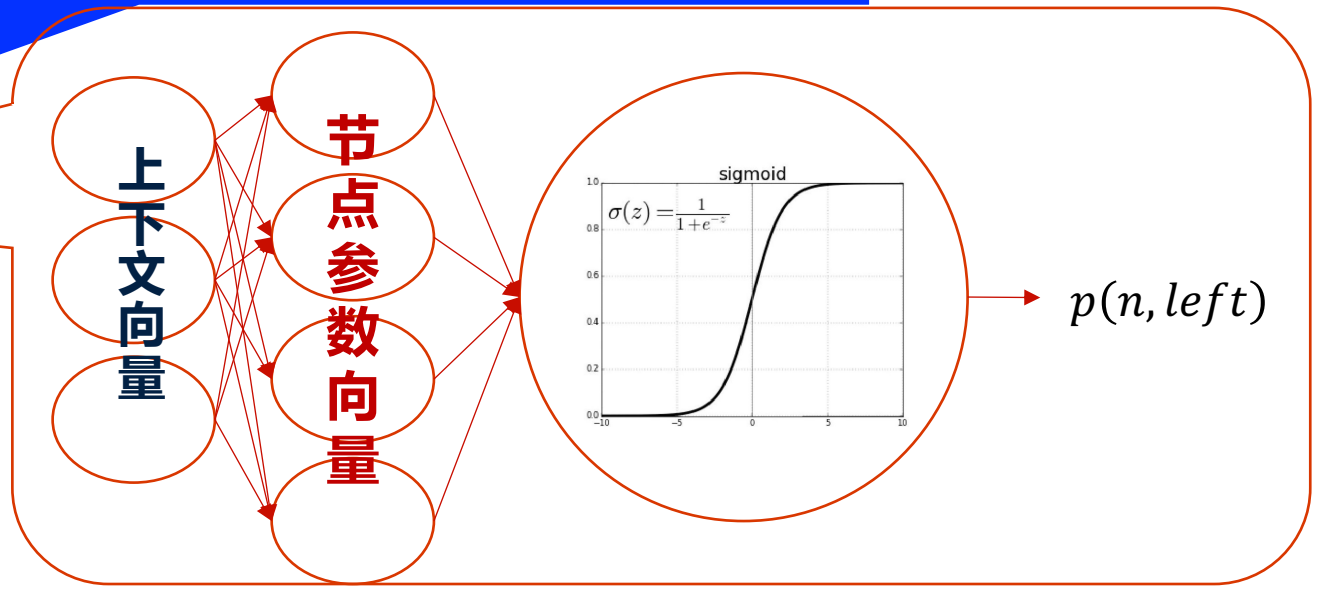
高频词被放置在树的浅层，低频词在深层

Hierarchical Softmax: $log_2(V)$

词频Huffman树

中间节点为sigmoid二分类器，需训练参数





$$p(n,\text{left}) = \sigma\left(\mathbf{v}_n'^{T} \cdot \mathbf{h}\right)$$

$$p(n,\text{right}) = 1 - \sigma\left(\mathbf{v}_n'^{T} \cdot \mathbf{h}\right) = \sigma\left(-\mathbf{v}_n'^{T} \cdot \mathbf{h}\right)$$

高频词被放置在树的浅层，低频词在深层

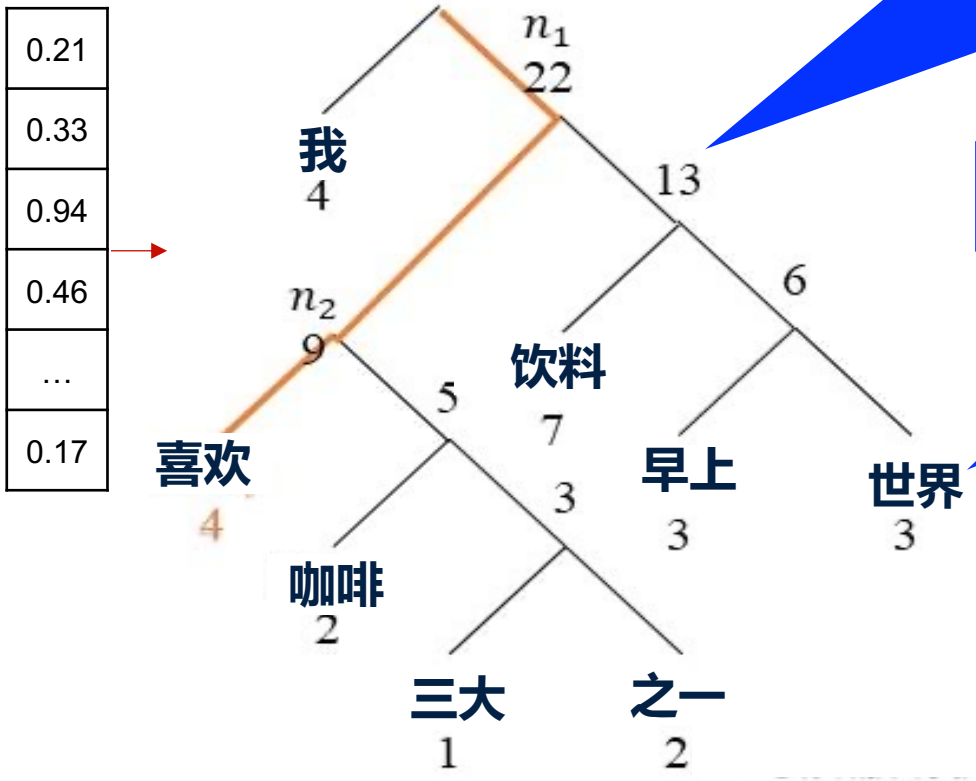
Hierarchical Softmax: $log_2(V)$

词频Huffman树

中间节点为sigmoid二分类器，需训练参数

$$p(n, \text{left}) = \sigma\left(\mathbf{v}_n'^{T} \cdot \mathbf{h}\right) \quad p(n, \text{right}) = 1 - \sigma\left(\mathbf{v}_n'^{T} \cdot \mathbf{h}\right) = \sigma\left(-\mathbf{v}_n'^{T} \cdot \mathbf{h}\right)$$

叶子节点为词，其概率为达到路径节点概率之积



高频词被放置在树的浅层，低频词在深层

$$p\,(\text{喜欢}|\text{我}, \text{早上}, \text{喝})$$

$$= p(n_0, right) \cdot p(n_1, left) \cdot p(n_1, left)$$

$$= \sigma\left(-v'^{T}_{n_0} \cdot vc\right)\sigma\left(v'^{T}_{n_1} \cdot vc\right)\sigma\left(v'^{T}_{n_2} \cdot vc\right)$$

$$v_c = \frac{1}{3}\left(v(\text{我}) + v(\text{早上}) + v(\text{喝})\right)$$

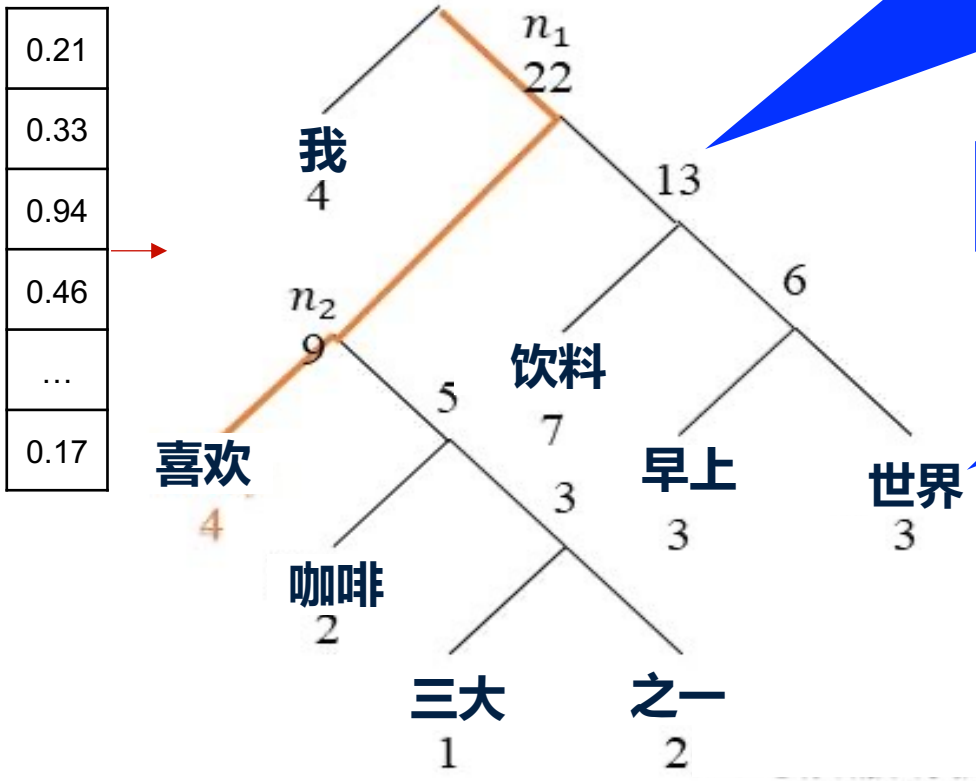
Hierarchical Softmax: $log_2(V)$

词频Huffman树

中间节点为sigmoid二分类器，需训练参数

$$p(n,\text{left}) = \sigma\left({\mathbf{v}'_n}^T \cdot \mathbf{h}\right) \quad p(n,\text{right}) = 1 - \sigma\left({\mathbf{v}'_n}^T \cdot \mathbf{h}\right) = \sigma\left(-{\mathbf{v}'_n}^T \cdot \mathbf{h}\right)$$

叶子节点为词，其概率为达到路径节点概率之积



高频词被放置在树的浅层，低频词在深层

优化目标

$$p(w = w_O) = \prod_{j=1}^{L(w)-1} \sigma\left(\llbracket n(w,j+1) = \text{ch}(n(w,j)) \rrbracket \cdot {\mathbf{v}'_{n(w,j)}}^T \mathbf{h}\right)$$

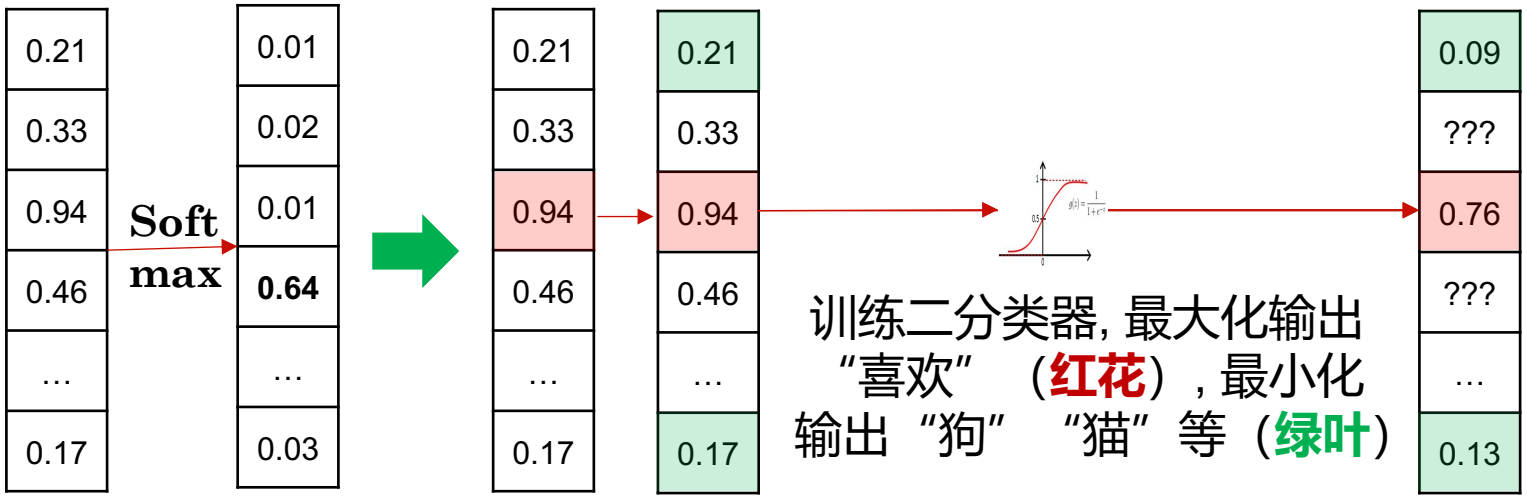
损失函数

$$E = -\log p(w = w_O | w_I) = -\sum_{j=1}^{L(w)-1} \log \sigma(\llbracket \cdot \rrbracket {\mathbf{v}'_j}^T \mathbf{h})$$

简化书写

另一种Softmax优化方法：Negative Sampling

采样原则：分母不用全部词求和了，就选几个代表



$$E = -\log \sigma(\mathbf{v}'_{w_O}{}^T \mathbf{h}) - \sum_{w_j \in \mathcal{W}_{\text{neg}}} \log \sigma(-\mathbf{v}'_{w_j}{}^T \mathbf{h})$$

正样本（目标词）　上下文词向量　负样本（绿叶词）

目录

GloVe (Global Vectors for Word Representation)

[PDF] Glove: Global vectors for word representation

J Pennington, R Socher… - Proceedings of the 2014 …, 2014 - aclanthology.org

… GloVe, … GloVe model outperforms all other methods on all evaluation metrics, except for the CoNLL test set, on which the HPCA method does slightly better. We conclude that the GloVe …

☆ 保存 99 引用 被引用次数：49772 相关文章 所有 25 个版本 »

特性	GloVe	Word2Vec（CBOW/Skip-gram）
核心依据	全局共现统计 + 局部上下文	仅局部上下文（窗口内）
训练目标	拟合共现频率的对数关系	预测任务（上下文→中心词 / 中心词→上下文）
低频词效果	优（全局统计补偿低频词样本不足）	一般（仅依赖局部窗口样本）
语义类比能力	强（如 “国王 - 男人 + 女人 = 王后”）	中（依赖局部语义关联）
训练速度	中等（需先构建共现矩阵）	快（直接处理分词语料）
内存占用	较高（共现矩阵可能达 V×V 规模）	较低（无需存储全局统计）

主要思路

- Word2Vec用了词的局部上下文，但没考虑词在语料中的全局信息
- GloVe使用全局信息的方法：看**词对共现概率的比值**
  - “（水丨冰）”和“（水丨蒸汽）”：共现概率比较像
  - “（固体丨冰）”比“（固体丨蒸汽）”：共现概率大得多
  - “（气体丨冰）”比“（气体丨蒸汽）”：共现概率小得多
  - “（时尚丨冰）”和“（时尚丨蒸汽）”：共现概率比较像

（冰丨水）：
冰在水上下文中的条件概率

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

主要思路

- GloVe使用全局信息的方法：看**词对共现概率的比值**
  - “（水丨冰）”和“（水丨蒸汽）”：共现概率比较像
  - “（固体丨冰）”比“（固体丨蒸汽）”：共现概率大得多

词对共现概率 $P_{ij}=P(w_j|w_i)=\frac{X_{ij}}{X_i}$

Probability and Ratio	$k=solid$	$k=gas$	$k=water$	$k=fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

主要思路

- GloVe使用全局信息的方法：看**词对共现概率的比值**
  - “（水丨冰）”和“（水丨蒸汽）”：共现概率比较像
  - “（固体丨冰）”比“（固体丨蒸汽）”：共现概率大得多

词对共现概率的比值 $\frac{P_{ik}}{P_{jk}}$

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

主要思路

假设有一个以词向量为输入的函数F, 可以计算这个比值

$$F(w_i, w_j, w_k) = \frac{P_{ik}}{P_{jk}}$$

$$F(w_i - w_j, w_k) = \frac{P_{ik}}{P_{jk}}$$

$$F\left((w_i - w_j)^T, w_k\right) = \frac{P_{ik}}{P_{jk}}$$

主要思路

假设有一个以词向量为输入的函数F, 可以计算这个比值

$$F(w_i, w_j, w_k) = \frac{P_{ik}}{P_{jk}}$$

$$F(w_i - w_j, w_k) = \frac{P_{ik}}{P_{jk}}$$

$$F\left((w_i - w_j)^T, w_k\right) = F(w_i^T w_k - w_j^T w_k) = \frac{P_{ik}}{P_{jk}}$$

什么函数是减法变除法?

主要思路

假设有一个以词向量为输入的函数F, 可以计算这个比值

$$e^{x-y} = \frac{e^x}{e^y} \Rightarrow F(\cdot) = e^{(\cdot)}$$

$$F\left(w_i^T w_k - w_j^T w_k\right) = \frac{e^{w_i^T w_k}}{e^{w_j^T w_k}} = \frac{P_{ik}}{P_{jk}}$$

$$w_i^T w_k = log(P_{ik}) = log\left(\frac{X_{ik}}{X_i}\right) = log(X_{ik}) - log(X_i)$$

目标函数

$$w_i^T w_k = log(X_{ik}) - log(X_i)$$

相等　　相等　　不等

$$w_k^T w_i = log(X_{ki}) - log(X_k)$$

可训练参数

$$w_i^T w_k = log(X_{ik}) - b_i - bk$$

$$w_k^T w_i = log(X_{ki}) - b_k - bi$$

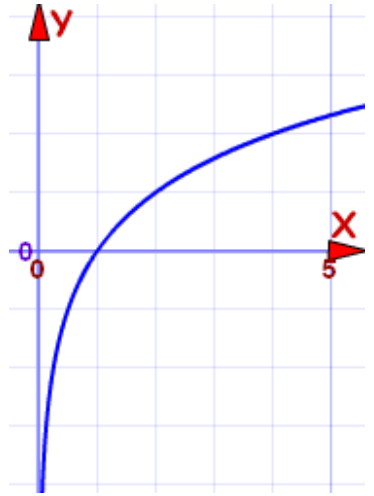
目标函数

$$w_k^T w_i = log(X_{ki}) - b_k - bi$$

$$argmin\left(\sum_{i,j=1}^{V}\left(w_i^T w_j + bi + bj - log(X_{ij})\right)^2\right)$$

目标函数

$$argmin\left(\sum_{i,j=1}^{V}\left(w_i^T w_j + bi\ + bj - log(X_{ij})\right)^2\right)$$



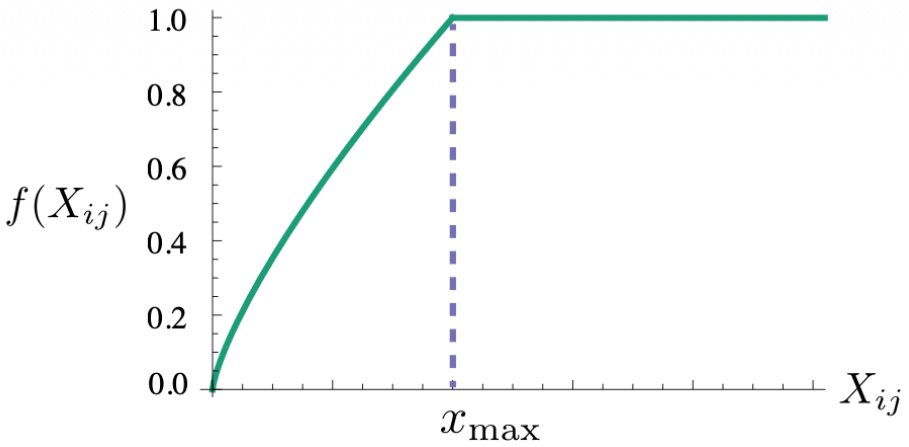
$$0$$

$$argmin\left(\sum_{i,j=1}^{V} f(X_{ij})\left(w_i^T w_j + bi\ + bj - log(X_{ij})\right)^2\right)$$

目标函数

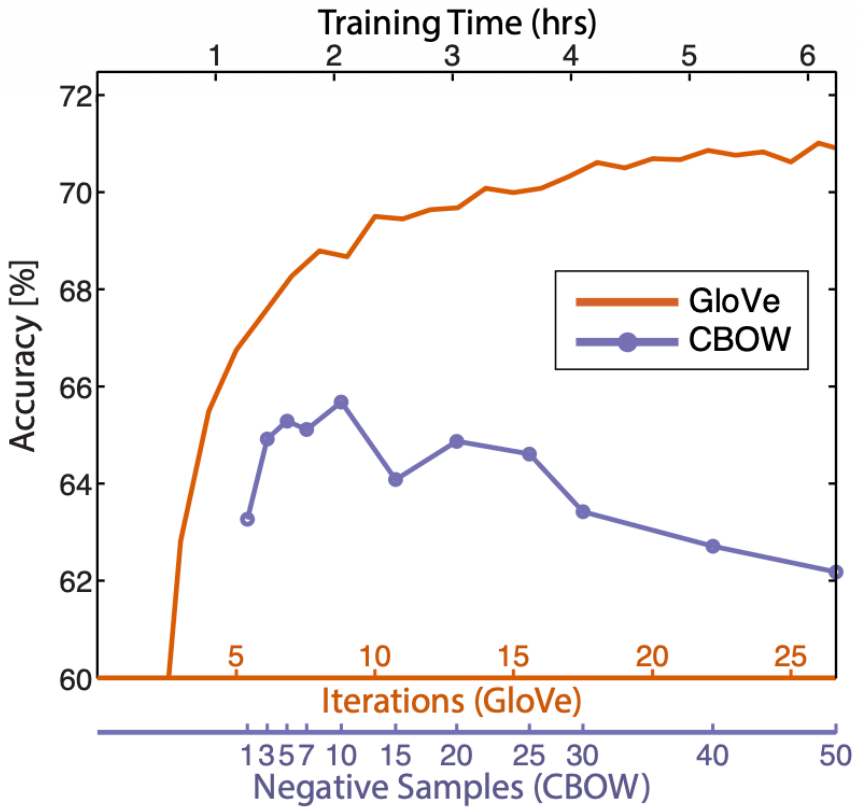
$$argmin\left(\sum_{i,j=1}^{V} f(X_{ij})\left(w_i^T w_j + bi + bj - log(X_{ij})\right)^2\right)$$

$$f(X_{ij}) = \begin{cases} (x/x_{max})^{\alpha}, if\ x < x_{max} \\ 1, \quad otherwise \end{cases}$$

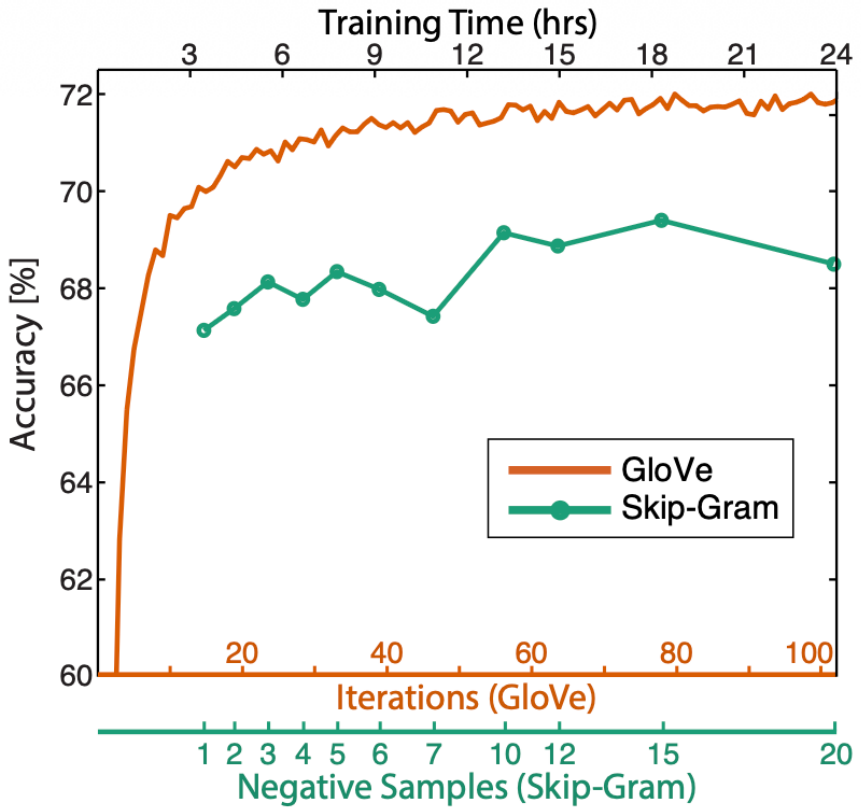


$$f(0) = 0$$

GloVe与Word2Vec对比



(a) GloVe vs CBOW



(b) GloVe vs Skip-Gram

下载地址

http://nlp.stanford.edu/projects/glove/

GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning

Introduction

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

Getting started (Code download)

- Download the latest latest code (licensed under the Apache License, Version 2.0).
  Look for "Clone or download"
- Unpack the files: unzip master.zip
- Compile the source: cd GloVe-master && make
- Run the demo script: ./demo.sh
- Consult the included README or Training_README for further usage details, or ask a question
- For more information on the data or training hyperparameters used for the 2024 vectors, please see the report

Download pre-trained word vectors

- Pre-trained word vectors. This data is made available under the Public Domain Dedication and License v1.0 whose full text can be found at: http://www.opendatacommons.org/licenses/pddl/1.0/.
  - **NEW!!** 2024 Dolma (220B tokens, 1.2M vocab, uncased, 300d vectors, 1.6 GB download): glove.2024.dolma.300d.zip
  - **NEW!!** 2024 Wikipedia + Gigaword 5 (11.9B tokens, 1.2M vocab, uncased, 300d vectors, 1.6 GB download):

问题

- 一词多义怎么办?

提示：如何使用自注意力?

1. 拼接：

2. 相加：

那时候， 柔嘉在家里等鸿渐回家来吃晚饭， 希望他会跟妯娌和好， 到她厂里做事

中国科学院大学
University of Chinese Academy of Sciences

目录

什么是词

- 词：自然语言中能够独立运用的最小单位
- 词法分析：
  - 形态还原：如英语、德语、俄语等，词形态变化表示语法关系
    - was → is
  - 分词：如汉语
    - 门/ 把手/ 弄/ 坏了/
    - 门把手/弄/ 坏了/
  - 黏着语：如：日语等，分词＋形态还原

传统分词的挑战1

- 组合型歧义
  - 门/ 把手/ 弄/ 坏了/
  - 门/ 把/ 手/ 弄/ 坏了/
- 交集型歧义
  - 中国/ 人为/ 了/ 实现/ 自己/ 的/ 梦想
  - 中国人/ 为了/ 实现/ 自己/ 的/ 梦想
  - 中/ 国人/ 为了/ 实现/ 自己/ 的/ 梦想

中文分词歧义出现频度约为1.2次/100字
交集型歧义与组合型切分歧义的出现比例约为12:1

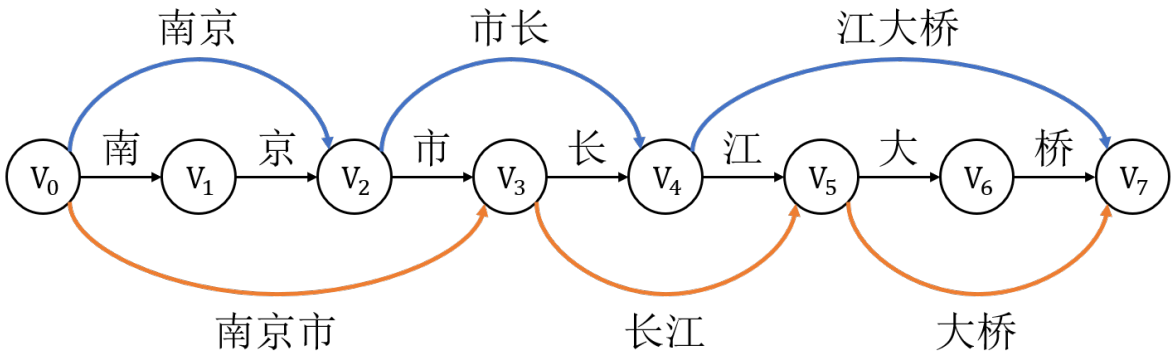
传统分词的方法一览

- 以词典为基础，使用匹配、统计、机器学习方法切分

输入文本：南京市长江大桥

给定词典：南、京、市、长、江、大、桥、南京、市长、长江、大桥、南京市、江大桥（人名）

词图：



最短路径 1：南京 / 市长 / 江大桥

最短路径 2：南京市 / 长江 / 大桥

传统分词的挑战2

- 未登录词（就是没有录入到词典中的词）
  - 新出现的词汇、术语、个别俗语等，例如：**词元**（Token）
  - 人名、地名、组织机构名等，例如：詹姆士
- **未登录词在中文中非常常见**

大模型时代的分词

- **Token（词元）**
  - 在大模型中，词逐渐变为Token
  - Token不是传统意义上的词，更像是一个连续文本中的“块”
    - 就像我们读俄文一样，读几遍可能就归纳出了一些连续的“块”

Token：搞区块链叫它“代币”，做网络安全的叫它“令牌”，编译器开发者叫它“标记”

3月23日，在中国发展高层论坛2026年年会上，国家数据局局长刘烈宏正式明确，AI领域核心术语Token的官方中文名为“词元”，并披露我国日均词元调用量已突破140万亿，较2024年初增长超1000倍。这一官宣终结了Token多年来“无统一中文名”的尴尬，也标志着我国AI产业发展进入规范化、标准化新阶段，相关消息与数据均

子词粒度的Tokenization

3.subword粒度Tokenization

subword粒度Tokenization介于词粒度和字符粒度之间，它将文本分割成介于单词和字符之间的子词（subwords）作为token。常见的subword Tokenization方法包括Byte Pair Encoding (BPE)、WordPiece等。这些方法通过统计文本数据中的子串频率，自动生成一种分词词典，能够有效应对未登录词（OOV）问题，同时保持一定的语义完整性。

```
1.  helloworld
```
复制

假设经过BPE算法训练后，生成的子词词典包含以下条目：

```
1.  h, e, l, o, w, r, d, hel, low, wor, orld
```
复制

子词粒度Tokenized结果：

```
1.  ['hel', 'low', 'orld']
```
复制

这里，“helloworld”被切分为三个子词“hel”，“low”，“orld”，这些都是词典中出现过的高频子串组合。这种切分方式既能处理未知词汇（如“helloworld”并非标准英语单词），又保留了一定的语义信息（子词组合起来能还原原始单词）。

1. Byte Pair Encoding (BPE)

- 字节对编码，基本思想：
  - 构建词表：从字符开始，**迭代合并**频率最高的对，形成词表

$$V_{t+1} = V_t \cup \max\_\text{freq}(sw_i, sw_{i+1})$$

  - 分词：用句子中**最长的子词先切分**，再用次长一步步切完
- 用于GPT-2等预训练模型，是应用最广的分词方法之一

1. Byte Pair Encoding (BPE)

- 构建词表：从字符开始，**迭代合并**频率最高的对，形成词表

$$V_{t+1} = V_t \cup \max\_\text{freq}(sw_i, sw_{i+1})$$

语料

- 中 国 地 理：10次
- 中 国 科 学 院：12次
- 科 学 院 南 路：8次

{中，国，地，理，科，学，院，南，路}

+中国:22次

1. Byte Pair Encoding (BPE)

- 构建词表：从字符开始，**迭代合并**频率最高的对，形成词表

$$V_{t+1} = V_t \cup \max\_\text{freq}(sw_i, sw_{i+1})$$

语料

- 中国 地 理：10次
- 中国 科 学 院：12次
- 科 学 院 南 路：8次

{中，国，地，理，科，学，院，南，路，中国}

+科学:20次

1. Byte Pair Encoding (BPE)

- 构建词表：从字符开始，**迭代合并**频率最高的对，形成词表

$$V_{t+1} = V_t \cup \max\_\text{freq}(sw_i, sw_{i+1})$$

语料

- 中国 地 理：10次
- 中国 科学 院：12次
- 科学 院 南 路：8次

{中，国，地，理，科，学，院，南，路，中国，科学}

+科学院:20次

1. Byte Pair Encoding (BPE)

- 构建词表：从字符开始，**迭代合并**频率最高的对，形成词表

$$V_{t+1} = V_t \cup \max\_\text{freq}(sw_i, sw_{i+1})$$

语料

- 中国 地 理：10次
- 中国 科学院：12次
- 科学院 南 路：8次

{中，国，地，理，科，学，院，南，路，
中国，科学，科学院}

1. Byte Pair Encoding (BPE)

- 构建词表：从字符开始，**迭代合并**频率最高的对，形成词表

$$V_{t+1} = V_t \cup \max\_\text{freq}(sw_i, sw_{i+1})$$

语料

- 中国 地理：10次
- 中国 科学院：12次
- 科学院 南路：8次

{中，国，地，理，科，学，院，南，路，
中国，科学，科学院，地理，南路}

1. Byte Pair Encoding (BPE)

- 分词：用句子中**最长的子词先切分**，再用次长一步步切完

{中，国，地，理，科，学，院，南，路，中国，科学，科学院，地理，南路}

中国科学院地理所 → 中国 科学院 地理 所

1. Byte Pair Encoding (BPE)

- 优点：
  - 解决了未登录词问题
  - 通过控制词汇表大小，可以控制分词粒度
  - 性能较高

性能对比（10GB英文文本）：

分词方法	词表大小	OOV率	序列长度	处理速度
单词级	500K+	15.2%	1.0x	1.0x
字符级	256	0%	5.8x	0.3x
BPE	50K	0.3%	1.2x	1.5x

2. WordPiece

- 思路：与BPE类似，差别在于BPE按频率来选择合并的token对，而WordPiece按token间的互信息来进行合并
- 应用模型：BERT

3. ULM（Unigram language model)

- 思路：通过unigram 语言模型计算删除不同subword造成的损失, 保留重要性较高的subword
- 应用模型： XLNet/ALBERT/Marian/T5

3. ULM（Unigram language model)

- 步骤：
  - 准备基础词表：初始化一个很大的词表，比如所有字符+高频ngram，也可以通过BPE算法初始化；
  - 针对当前词表，用EM算法估计每个子词在语料上的概率；
  - 计算删除每个subword后对总loss的影响，作为该subword的loss；
  - 将子词按照loss大小进行排序，保留前x%的子词；注意，单字符不能被丢弃，以免OOV；
  - 重复步骤2到4，直到词表大小减少到设定值

SentencePiece分词工具包

- 谷歌推出的子词开源工具包
  - 支持BPE、ULM子词算法，也支持char，word分词
  - 多语言：以unicode方式编码字符，将所有的输入都转化为unicode字符
  - 编解码的可逆性：显式地将空白作为基本标记来处理，用一个元符号“_”（U+2581）转义空白，实现简单且可逆编解码
  - Fast and lightweight

https://github.com/google/sentencepiece

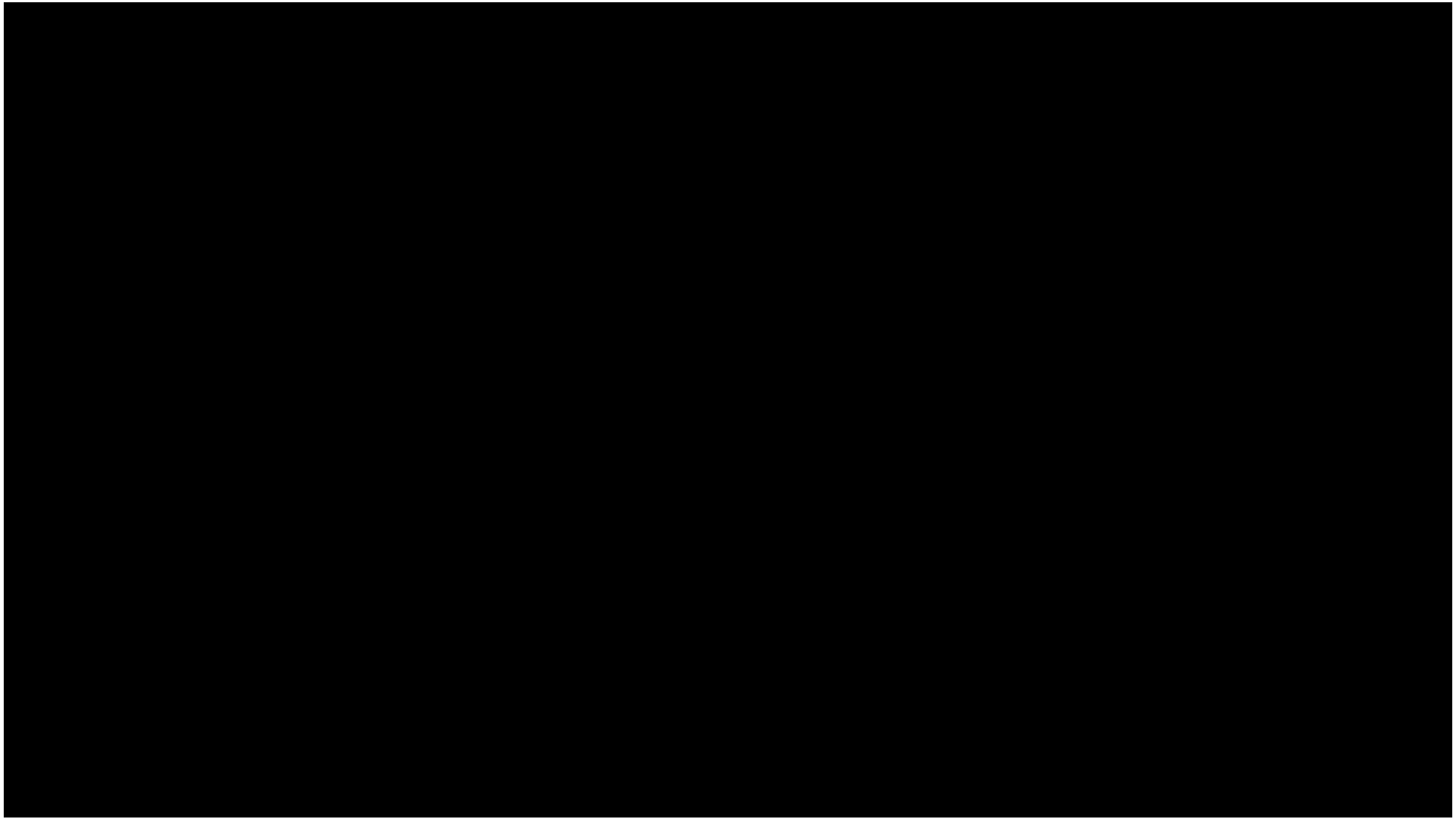
本节复习

- Embedding
- Word2Vec: CBOW, Skip-Gram
- Glove: ice, cream …
- BPE

参考文献

- Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." nips2013
- Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." EMNLP 2014.

分布式表示是当前NLP的基础



致 谢

- 胡玥、曹亚男、方芳：国科大《自然语言处理基础》
- 曹亚男、任昱冰：国科大《深度学习与自然语言处理概述》


THANKS
https://ictkc.github.io/teaching/2026spring-nlp