

Natural Language Processing
#L6
词义分析

■什么是意义？

- 语言中争议最多的术语之一
- C.K.Ogden 和 I.A.Richards: 意义之意义(1923): 22种
 - 符号使用者实际所指的事物
 - 符号使用者应指的事物
 - 符号使用者自认为所指的事物
 - 符号解释者所指的事物
 - 词典上加在一个词之后的其他词
 - 一种内在的特征
 - 一个体系中任何一个事物的部位
 - ...

■什么是意义？

七种不同的意义 (J. N. Leech,语义学(1981) 第二版)

	理性意义	关于逻辑、认知或外延内容的意义
联想意义	内涵意义	通过语言所指事物来传递的意义
	社会意义	关于语言运用的社会环境的意义
	情感意义	关于讲话人/写文章的人的感情和态度的意义
	反映意义	通过与同一个词语的另一意义的联想来传递的意义
	搭配意义	通过经常与另一个词同时出现的词的联想来传递的意义
	主题意义	组织信息的方式(语序、强调手段)所传递的意义

■词义：词的内容，可以分为概概念义、色彩义

■概概念义——是词义中反映**客观**事物自身的那部分内容，又称为客观义、理据义或指称义。

■例子：太阳、月亮、思想、观点等词中所反映的客观对象

■色彩义——着重反映了人们的**主观**认识。色彩义有：感情义、雅俗义、古今义、地域义、社区义、时间义、语体义、修辞义等。

■概概念义相近，色彩义不同的例子：

■智囊—狗头军师：感情义不同：前者褒义，后者贬义

■走：古今义不同：古义为跑

■好看—美丽：语体义不同：前者口语色彩，后者书面色彩

■义素(Sememe)：构成义位的最小词义单位(相对性)。

■义位(glossememe)：能独立运用的最小词义单位。

■义素分析法：

- 确定对比分析的义位，通常是同一语义场内的一些义位。
- 寻找义位之间的共性特征和区别性特征。
- 将寻找出的各种义素用结构式描述出来。

	Human	Male/female	Adult/young
Woman	+	-	+
Man	+	+	+
Girl	+	-	-
Boy	+	+	-

[ZhangWY2001]张万有，义素分析略说，语言教学与研究，2001,(01)



- 义素(Sememe)：最小词义单位。
- 义位(glosseme)：能独立运用的最小词义单位。
- 义项(sense)：词典中对词义进行符合语言发展规律和规则的分项解说的系列或其中的每个分项。[TangCQ1985]
 - 可以是一个义项对应一个义位，也可以是一个义项反映两个甚至多个相邻近的义位。
 - 义项的分合体现了对义位的不同组织方式，相比而言，义位是客观的，义项是主观的。

- 提供词义的词典：为人获取词义而用的

- 打：(部分选自现代汉语词典)
 - [s1]12个构成的一组，如一打铅笔
 - [s2]击：身世浮沉雨打萍
 - [s3]斗殴：打架
 - [s4]进攻：围点打援
 - [s5]射击：打枪
 - ...
 - [s20]表示身体上的某些动作：打哈欠、打冷战
 - [s21]从，自：打哪里来？

- 目前的计算机使用起来还比较困难

- 问题：如何让计算机获取其能够使用的词义？

Sense 【义项】

Meaning 【意义】

Semantics 【语义】

■问题：如何让计算机获取其能够使用的词义？

■从获取渠道上看：

■人工(专家)总结

■自动从语料中获取

■从获取的词义的表示方法来看：

■基于符号的表示方法

■基于数值(向量)的表示方法

■人工总结的多是符号表示，而自动获取可能是符号表示，更多是数值表示。

■从借鉴词典的词义定义开始：

■空气：(在线汉语词典<http://xh.5156edu.com/>)

■[1]构成地球周围大气的气体。(义项1)

■.....

■气体：

■[1]没有一定的形状和体积、能充满任何容器的物质。

■.....

■词义表达的方式：

■用有限的基本词

■或者用与其他词(不一定是基本词)的关系

- 基于一个基本词(概念)集合来表示其他所有词的语义
- 一个途径：基本词是义素(语义特征\义原)
- 主要问题：如何找到基本词？
- 主要方法：
 - 专家定义
 - 自动发现



■ 基于专家定义的方法，例1：知网(HowNet)

■ 知网是一个基于义原的常识知识库，基于义原揭示概念与概念之间以及概念所具有的属性之间的关系。

■ 义原定义过程

■ 首先对汉字（单纯词）进行考察和分析，获取一些义原

■ 然后用这些义原作为标注集去标注多音节的词，当发现这些义原不满足要求时，便进行调整或扩充。

■ 这样最终形成了2000多个义原的标注集以及由它们标注的10万个中文/英文词或短语。

■ HowNet: <http://www.keenage.com/>

■ OpenHowNet: <https://openhownet.thunlp.org/home>

■打

- 打1: DEF=buy|买 (义原)
 - E_C=~酱油, ~张票, ~饭, 去~瓶酒, 醋~来了
- 打2: DEF=weave|辫编 (义原)
 - E_C=~毛衣, ~毛裤, ~双毛袜子, ~草鞋, ~一条围巾,
~麻绳, ~条辫子
- 打: 我女儿给我打的那副手套哪去了
 - d(手套,酱油) vs d(手套,毛衣) ?
- 词的关系归结到其义原间的关系 [LiuQ2002]

■ 基于专家定义的方法，例2：词典

■ 很多词典是基于专家确定的一些基本词而构建出来的，如Oxford学生词典：专家确定2000基础词，其他词均由基础词定义，因此从这些词典中可以获取方式

■ 可能的方式：

■ 词：**语言**里最小的、可以自由运用的单位

■ 语言：人类最重要的**交际**工具……。

■ 交际：往来**应酬**

■ 应酬：交际**往来**

■ 往来：**交往**，交际

■ 交往：互相**来往**

■ 来往：交际**往来**

■ 来往=交往 为一个基本词

- 前面均为基于专家的人工方法
 - 主观性，不一致，费时费力
 - 新词不断出现、词义逐渐变化，
- 近年来出现了一些基于语料来自动获取基本词的研究(Sememe Prediction)
 - [XieRB2017IJCAI] Ruobing Xie, et al. Lexical sememe prediction via word embeddings and matrix factorization. IJCAI2017.
 - 基于词向量的相似性预测一个词是否包含某个义原。
 - Wei Li, et al. Sememe Prediction: Learning Semantic Knowledge from Unstructured Textual Wiki Descriptions. arXiv 2018.
 - Fanchao Qi, et al. Cross-lingual Lexical Sememe Prediction. EMNLP 2018
 - Huiming Jin et al. Incorporating Chinese Characters of Words for Lexical Sememe Prediction. ACL 2018.

■义素的心理学支持: 概念的特征表说

■概念由两部分组成:

- 一些特征: 一个特征就是一个义素
- 一些规则: 如何结合不同特征的规则

■例子:

■概念“红圆”

- 特征集 = {“红”, “圆”}
- 规则: 特征”and”操作

- 义素与语言使用有关联
- 义素可以更细致地约束词的使用
 - 如果某个词有一个义素是“液体”，则
 - 它很可能与”泼”、“喝”一起使用 ...
 - 而与“吃”、“锯”一起使用的话可能有问题 ...
- → 反之，也可作为义素抽取的一种途径
 - 如果一个词和”泼”、“喝”一起使用，则可能该词具有“液体”这个语义特征

■ 义素分析的问题

- 总共有多少义素？

- 多少义素足够用且数量最少？

- 如何获得义素？

- 专家人工定义？机器自动获取？

- 人工概念相对容易、自然概念比较难统一。

- 人工定义难以快速跟上新词以及词义的发展

- 机器自动抽取在研究中、提出新义原还没有

■词间关系，主要是词义之间的关系，当词多个义位时，看其中一个义位间的关系。所以更准确地说两个词之间的关系是指这两个词的某两个义位之间的关系。

■上下义位关系(Hyponymy)

■全体-成员关系(Ensemble - Member)

■整体-部分关系(Whole-Part)

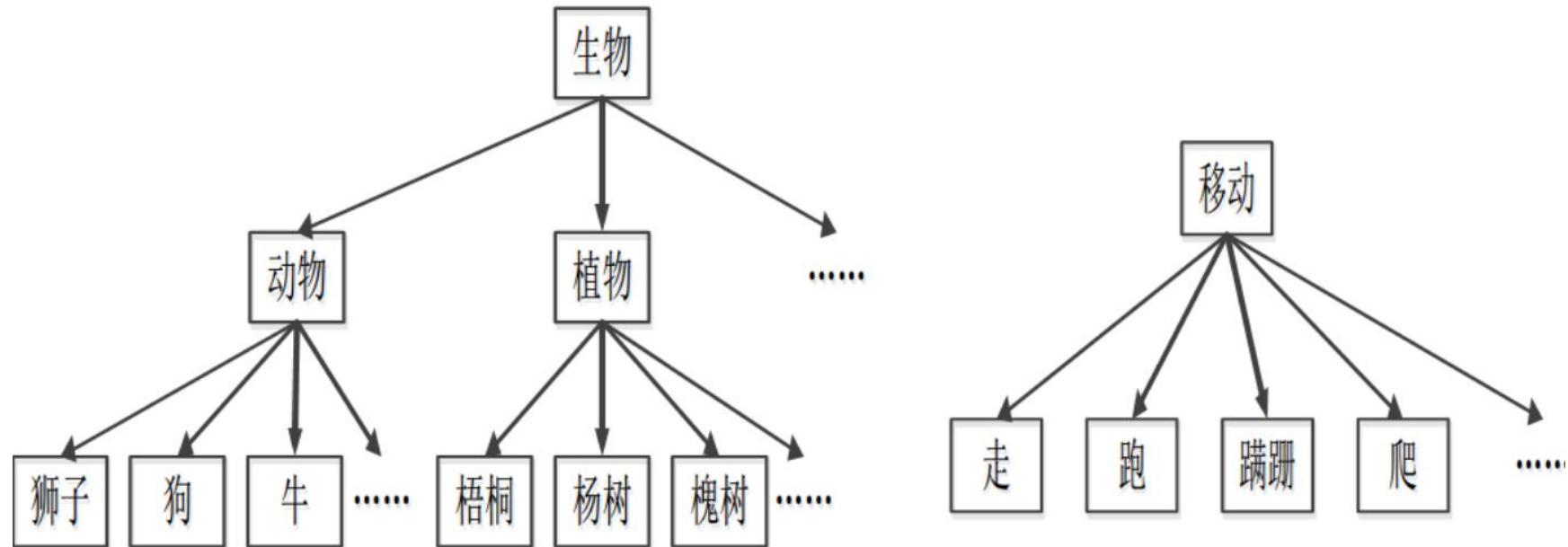
■同义关系(Synonymy)

■.....

上下义位关系：上义位，下义位

- 上位(Superordinate上位词)：从特殊概念到一般概念 (IS-A)
 - 哺乳动物→动物
 - 层次结构：哺乳动物→动物→生物→物...
- 下位(Subordinate下位词)：从一般概念到特殊概念 (Include)
 - 动物→哺乳动物
 - 层次结构：动物→哺乳动物→老虎→东北虎...

名词、动词等都可以有上下位概念结构



■全体-成员关系

■从全体到成员关系 (Has-Member) : :

■学会→会员

■从成员到全体: :

■会员→学会

■整体-部分关系(Whole-Part)

■从整体到部分: Part Meronym(Has-Part):

■桌子→腿

■从部分到整体:Part Holonym(Part-Of):

■腿→桌子

两个词(基于音、义位)之间的关系：

- 同音词 (Homonyms)
- 同形(同音异义) 词(Homographs)
- 同形异音异义词(Heteronyms)
- 近义词 (Synonyms)
- 反义词(Antonyms)
- 换喻(Metonyms)
- ...

- 同音词 (Homonyms)
 - 读音相同的两个词
 - tale and tail ; 红 和 洪
- 同形(同音)异义词 (Homographs)
 - 写法相同的两个词
 - dove (鸽子, dive的过去时) ; 省(行政单位, 节约)
- 同形异音异义词(Heteronyms)
 - 看(kan4, kan1)

■近义词 (Synonyms): 具有相同或相近义位的不同词.

■Please do not annoy, torment, pester, plague, molest, worry, badger, harry, harass, heckle, persecute, irk, bullyrag, vex, disquiet, grate, beset, bother, tease, nettle, tantalize, or ruffle the animals. (San Diego Zoo Wild Animal Park) 22个动词

■打扰\折磨\纠缠\造成麻烦\骚扰\使烦恼\纠缠不休\一再骚扰\反复袭击\困扰\迫害\激怒\欺凌\使恼怒\使不安\使烦恼难受\困扰(围困)\使迷惑\戏弄、挑逗\惹恼\逗弄\使生气

■ 反义词(Antonyms): 不同的反

■ 互补

生/死(Alive/dead), 出席/缺席(present/absent)...

■ 分级

■ 大/小(Big/small): ... 庞大-巨大-大-中-小-微小...

■ 关系

■ 给/收(Give/receive), 买/卖(buy/sell),...

■ 自反

■ Cleave(to split apart | to cling together), 无价?

■换喻 (Metonyms)

- 用一个对象来指称另一个对象
 - 从摇篮到坟墓：指称物的处所
 - 屋内空气紧张：气氛
- 用一个对象的属性或某个侧面来指称另一个对象
 - 红色政权：
- 用一个属性来指称一个对象
 - 赤色、黄色
- ...

■多义词(Polysemy)

■一个词有多个义项

■例如：他说：“她这个人真有意思(funny)”。她说：“他这个人怪有意思的(funny)”。于是人们以为他们有了意思(wish)，并让他向她意思意思(express)。他火了：“我根本没有那个意思(thought)”！她也生气了：“你们这么说是什么意思(intention)”？事后有人说：“真有意思(funny)”。也有人说：“真没意思(nonsense)”。

■ 区分

- 多义词(Polysemy) vs. 同形异义词(Homographs)
- 系统性(Systematically) vs. 偶然性(Occasionally)
- 不同义位间有系统性关联 vs. 没有关联

■ 例如：

- 省： [s1]节约; [s2]行政单位。
 - 二者无关
- 看： [s1]视; [s2]阅读; [s3]认为。
 - 三者相关且系统性相关 → 听、闻、触、嗅...

■ 虽然词可以有多个义项，但是在在一个信息足够充分的特定上下文中，只能取其中一个，让计算机在给定上下文时选择词的义项，是词义分析中的一个基本任务：词义消歧(Word Sense Disambiguation: WSD)

■ 两类具体任务：

- Lexical Sample任务
- All-words WSD任务

■ Lexical Sample任务：从词典里为给定上下文的某个特定词选择正确的义项

■ 例子：

■ 特定词： horse

■ 来自词典(WordNet)的horse的义项

■ Sense 1: horse, Equus caballus...

■ Sense 2: horse -- a padded gymnastic apparatus on legs...

■ Sense 3: cavalry, horse cavalry, horse...

■ Sense 4: sawhorse, horse, sawbuck, buck...

■ Sense 5: knight, horse ...

■ Sense 6: heroin, diacetyl morphine, H, horse, junk, scag, shit, smack...

■ 每个义项有若干标注样本

■ Sense 1: 1. He **ride** a horse. 2....

■

■ 基于这些样本为一个新句子中的该词选择合适的义项

■ All-words WSD任务：从词典里为文中所有目标多义词选择的义项

■ 例子：

■ 包含多个多义词的句子：

■ The art of **change-ringing** is peculiar to the **English** and, like **most English peculiarities** , **unintelligible** to the **rest of the world** . -- Dorothy L. Sayers , "The Nine Tailors " ASLACTON , **England** -- Of all **scenes** that **evoke rural England** , this is one of the **loveliest** : An **ancient stone church stands** amid the **fields** , the **sound of bells cascading** from its **tower** , **calling the faithful to evensong** . The **parishioners** of St. Michael and All Angels **stop to chat** at the **church door** , as **members here always have** .

■ 每个多义词都有来自词典(WordNet)的义项

■ 设计一个算法为其中的每个多义词选择合适的义项

词间相似性：比近义词更一般的词关系

■词间(义位)相似性

■两个词相似有很多层面：语义相似、用法相似、形态相似等，这里主要从语义的角度。

■ $\text{sim}(\text{猫}, \text{狗}) > \text{sim}(\text{猫}, \text{桌子})$ 、 $\text{sim}(\text{站}, \text{坐}) > \text{sim}(\text{站 vs 看待})$

■词间相似性具有很多重要的语言使用和理解价值：

■例如：

■如果： $\text{sim}(\text{猫}, \text{狗}) > \text{sim}(\text{猫}, \text{桌子})$

■那么：如果猫常和叫搭配使用，那么狗比桌子更可能和叫搭配使用。

■如何度量相似性？是一个重要的语言处理问题

词间(义位)相关性

■两组词

- 医生、病人、医院、急症
- 医生、毛巾、天空、经历

■前者构成一个语义场：这些词更可能共同出现在一个特定的场景中，具有相关性；

■相关性和相似性不同：

- 相关不一定相似：医生和急症相关，但不相似
- 相似一定相关：？

■词间相关性对于消歧等任务都具有重要价值

- 苹果：依据不同语义场中的相关词容易获得不同义项：“苹果的屏幕…”

■利用词间关系定义词义的例子：

■同义词词林(中文)

■梅家驹，竺一鸣，高蕴琦等编，同义词词林。上海辞书出版社，1983。

■哈工大社会计算与信息检索研究中心，同义词词林扩展版。

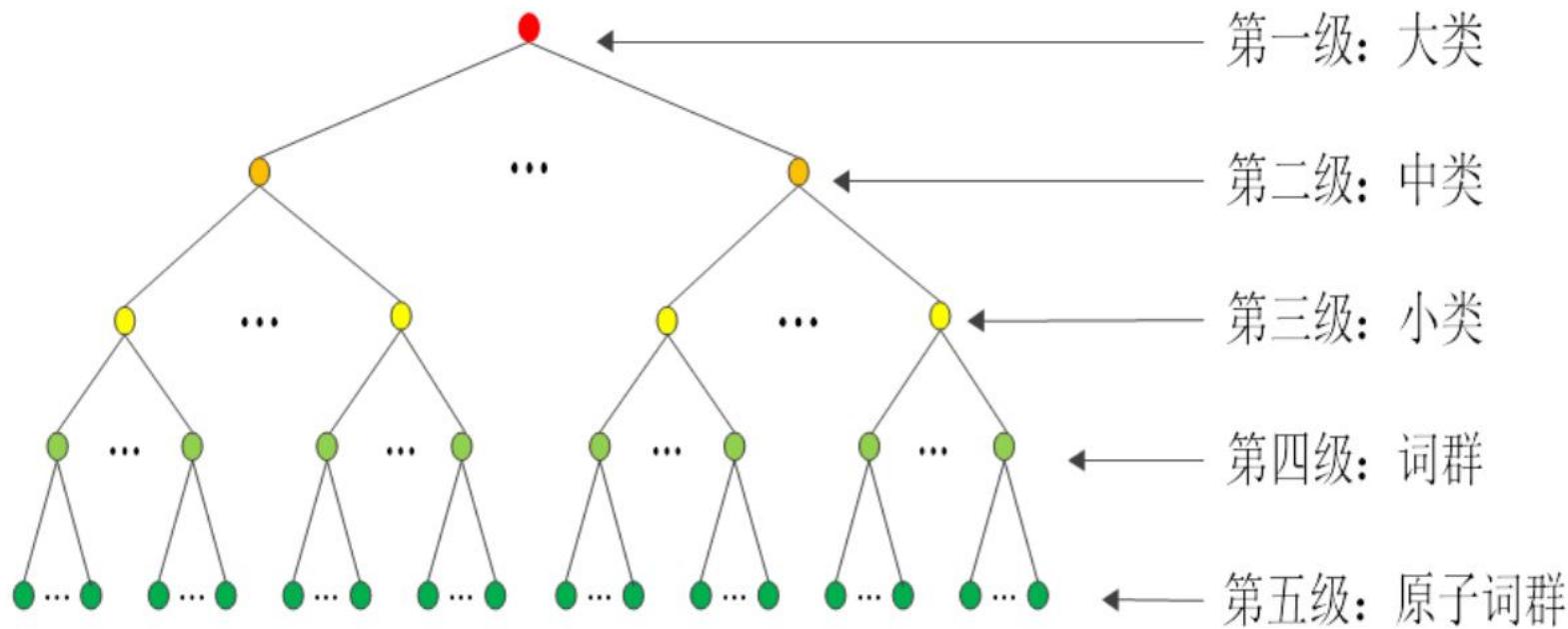
■WordNet(英文)

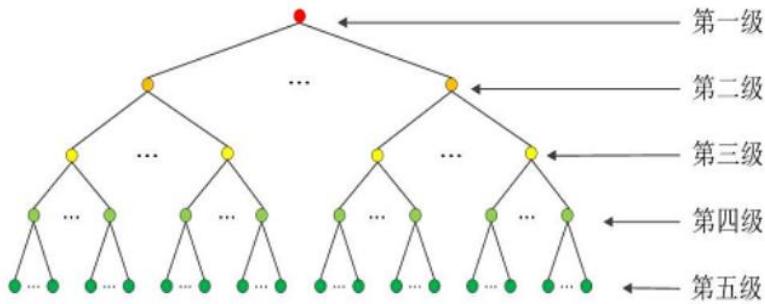
■George A. Miller

■<https://wordnet.princeton.edu/>



■ 同义词词林(中文)

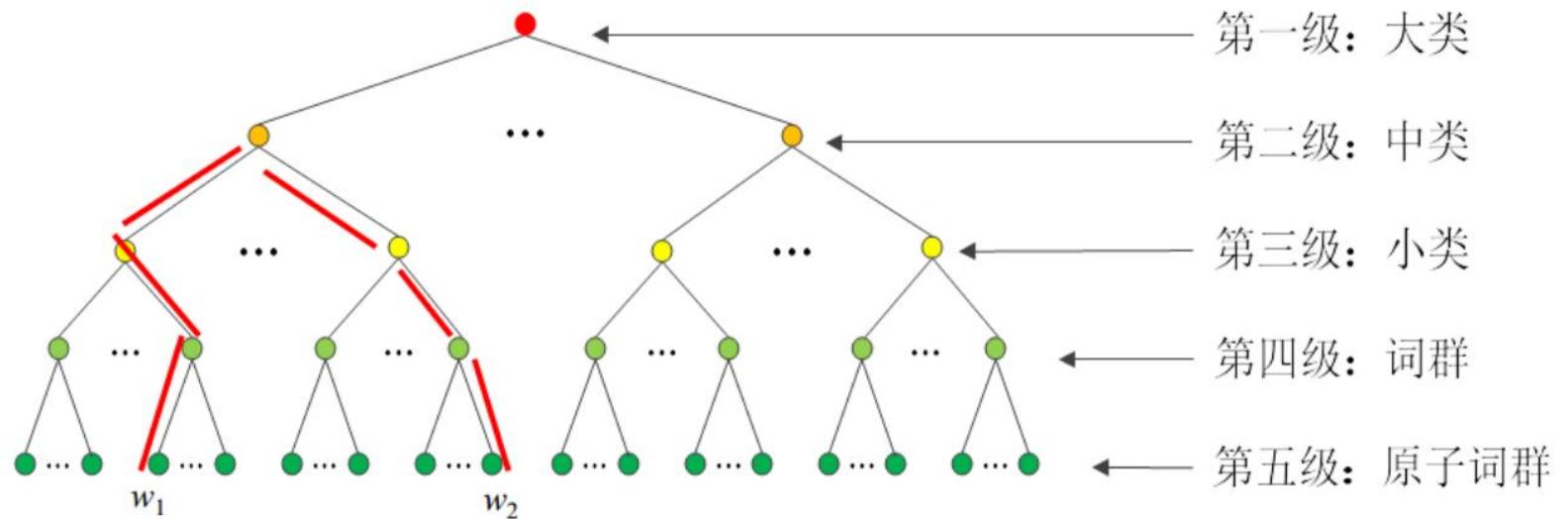




第一级	第二级	第三级	第四级	第五级
一个大 写字母	一个小 写字母	两位数字	一个大 写字母	两位数字
A 大类	a 中类	01 小类	A 词群	01 原子词群

- 原子词群 A a 01 A 01 = {人,士,人物,人士,人氏,人选}
- 原子词群 A a 01 A 02 = {人类,生人,全人类}
- 原子词群 A a 01 A 03 = {人手,人员,人口,人丁,口,食指}
- 原子词群 A a 01 A 04 = {劳力,劳动力,工作者}
- 词群 A a 01 A = {人,士,人物,人士,人氏,人选,人类,生人,全人类,.....}
- 小类 A a 01 = {...}
- 中类 A a = {...}
- 大类 A = {...}

- 同时也提供了计算词间关系的一些依据：
- 例如：词间的路径长度



■WordNet: 同义词集合(Synset)表示一个义项

■例如：

■synset: publish, print

■定义: put into print

■例子: These news should not be printed.

■synset: publish, write

■定义: have (one's written work) issued for publication

■例子: She published 25 books during her long career.

■synset: publish, bring out, put out, issue, release

■定义: prepare and issue for public distribution or sale

■例子: publish a magazine or newspaper.

George A. Miller, <https://wordnet.princeton.edu/>

■FrameNet：基于C. J. Fillmore的框架(frame)语义学

■Berkeley大学英语FrameNet (1997-)

■FrameNet：英语词汇数据库，人和机器均可读，基于词在真实文本中如何使用的标注样例而构建的

■<https://framenet.icsi.berkeley.edu/fndrupal/>

■山西大学汉语FrameNet (CFN) [YouLP2007]

■以框架语义学为理论基础、以真实语料为事实依据的汉语语义词典

■NTU Chinese FrameNet Lexicon (CFN-Lex) [YangTH2018]

■<http://www.nlg.csie.ntu.edu.tw/nlpresource/FrameNet/CFN-Lex/>

[YouLP2007]Liping You, Tao Liu, Kaiying Liu. Chinese FrameNet and OWL Representation, Sixth International Conference on Advanced Language Processing and Web Information Technology 2007. ALPIT 2007, pp. 140-145, 2007.

[YangTH2018]Tsung-Han Yang, Hen-Hsen Huang, An-Zi Yen, and Hsin-Hsi Chen. Transfer of Frames from English FrameNet to Construct Chinese FrameNet: A Bilingual Corpus-based Approach. LREC 2018.

■FrameNet：框架(frame)，例：

■框架：辨别 | Differentiation

语义结构

■核心框架元素：认知者，~~现象群(现象1、现象2、背景)~~

■非核心框架元素：工具、环境条件、程度、修饰、方法、结果

■词元：区分、区别、辨别、分别、分、别、辩明、明辨、分辨、辨、判别、甄别、鉴别、识别、辨认

相似、相关词

■ 上述基于词间关系表达词义的问题：

- 基于哪些词间关系？
- 词间关系多样、复杂、主观性高：近义词粒度
- 符号计算不直接：在目前的计算框架下不可直接计算，需要再设计量化的算法
- 例如：基于WordNet的词语相似性
 - 两个词之间的路径长度或进一步的改进度量

■ 基于某种更简单的关系、定量化

■ 数值表示的一些术语

- Word Distributed Representation(词分布式表示)
- Word Vector (词向量)
- Word Embedding (词嵌入)
- Continuous Space Word Representation (连续空间词表示)
-

■ 基于数值的词表示又进一步可以分为两个大类：

■ Point embedding:

■ 用N维实数向量(N维实数空间中的点)表示词，一个N维实向量表示一个词。

■ Gaussian embedding:

■ 用N维高斯分布来表示词，一个N维高斯分布表示一个词。

■ 目前研究和应用比较多的是前者，不过：

■ Point embedding的一些特性并不符合语言特性

■ 在后面的open question部分进一步阐述，并说明Gaussian embedding的一些特点和优势

- 获得embedding所使用的模型可以分为两大类：
 - Non-neural方法
 - Neural方法
- 进一步， Embeding在获取时可以结合多种信息/知识
 - 结合词形态信息、结合句法结构信息、结合外部知识
- 依据模型所利用的上下文以及embedding生成方案的不同， 可以分为：
 - 基于全局上下文的统计方法
 - 传统的先建立高维词向量，之后基于降维技术获得低维词向量
 - 基于局部上下文的预测方法
 - 直接获得低维词向量的word2vector(CBOW, SGNG)等
 - 这两种方法也可以进行结合，典型的代表是GloVe
- 本部分主要介绍这三种方法。

■ 基于向量的方法

■ 基于统计的高维向量及其降维

■ 基于预测的低维向量

- CBOW + Hierarchical SoftMax

- Skip-Gram + Negative Sampling

- 词向量评估

- 问题与发展

- 词向量应用

■ 分布式(distributional)词(义)表示

■ 基本思想

- J.R. Firth: 观其伴, 知其义(同现关系: 直接可观)
- 基于目标词上下文的词来定义目标词
- V 为词表 $W=(w_1, \dots, w_V)$ 的大小,
- 目标词 $w_t \in W$ 的向量表示 $x=(x_1, \dots, x_V)$
- x_i 的确定方法: 词 w_t 与 w_i 的同现情况
- 同现: 词 w_t 与 w_i 的在一个上下文中同时出现
- 上下文: 给定语料C和窗口K,
 - 布尔型: w_i 是否出现在 w_t 的窗口内 0/1
 - 频次型: w_i 出现在 w_t 的窗口内的次数 f

■示例：

■ $C=\{T_1, T_2, T_3\}$, $T_1=我\ 是\ 学生$; $T_2=你\ 是\ 学生$; $T_3=我\ 是\ 你\ 学生$ 。

■则：词表 $W=\{\text{我}, \text{你}, \text{是}, \text{学生}\}$

■上下文窗口 $K=5$ 时的表示：

	频率	布尔	独热(One-hot)
我	(0,1,2,2)	(0,1,1,1)	(1,0,0,0)
你	(1,0,2,2)	(1,0,1,1)	(0,1,0,0)
是	(2,2,0,3)	(1,1,0,1)	(0,0,1,0)
学生	(2,2,3,0)	(1,1,1,0)	(0,0,0,1)

■ 分布式(distributional)表示的优点

- 每个词对应一个向量，可以直接计算词间语义关系
- 每一维是有意义的(某个词)

■ 分布式(distributional)表示的问题

- 维数过高(维数=词表大小)，计算复杂度高
- 词间独立无关：同义词等在不同维

■ 将高维向量降维

■ 选择某些维(特征选择): tfidf...

■ 后面讲文本编码时会再提到

■ 压缩到特定维(特征抽取): LSI(Latent Semantic Index)...

■ LSI的核心是对频率矩阵进行SVD(Singular Value Decomposition)

■ SVD示例: 接上例, 先可由基于频率的表示得到word-word矩阵

	我	你	是	学生
我	0	1	2	2
你	1	0	2	2
是	2	2	0	3
学生	2	2	3	0

$$\rightarrow M = \begin{pmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 2 & 2 \\ 2 & 2 & 0 & 3 \\ 2 & 2 & 3 & 0 \end{pmatrix}$$

■ SVD: $M_{n \times m}$ 为 n 个 m 维向量, $\text{Rank}(M)=r$

■ $M_{n \times m} = U_{n \times n} S_{n \times m} V^T_{m \times m}$

■ 其中:

■ U 和 V 为酉矩阵(Unitary Matrix)

■ 即 $UU^T=I_n, VV^T=I_m$

■ $\lambda_1 \geq \lambda_2 \geq \dots \lambda_r > 0$

■ λ_r 为 MM^T 的特征值, M 的奇异值

$$S_{n \times m} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_r \\ \hline & 0 & & 0 \end{pmatrix}$$

■ → $M_{n \times m} = U_{n \times r} S_{r \times r} V^T_{r \times m}$

■ 其中: $S_{r \times r}$ 为对角阵, $S_{r \times r} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r)$,

■ 若取最大的 k 个奇异值, $k < r$, 即 $S_{k \times k} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$, 相应地 U 取前 k 列, 则: $U_{n \times k} S_{k \times k}$ 即为降到 k 维后的 n 个向量。

- 基于向量的方法
 - 基于统计的高维向量及其降维
 - 基于预测的低维向量
 - CBOW + Hierarchical SoftMax
 - Skip-Gram + Negative Sampling
 - 词向量评估
 - 问题与发展
 - 词向量应用

■基于神经网络的词向量学习

■早期：思想、小规模尝试

■[Hinton1986]G.E. Hinton, J.L. McClelland, D.E. Rumelhart.
Distributed representations. In: Parallel distributed processing:
Explorations in the microstructure of cognition. Volume 1:
Foundations, MIT Press, 1986.

■[Rumelhart1986]D. E. Rumelhart, G. E. Hinton, R. J. Williams.
Learning internal representations by back-propagating errors.
Nature, 323:533.536, 1986.

■[Elman1990]J. Elman. Finding Structure in Time. Cognitive
Science, 14, 179-211, 1990.

■.....

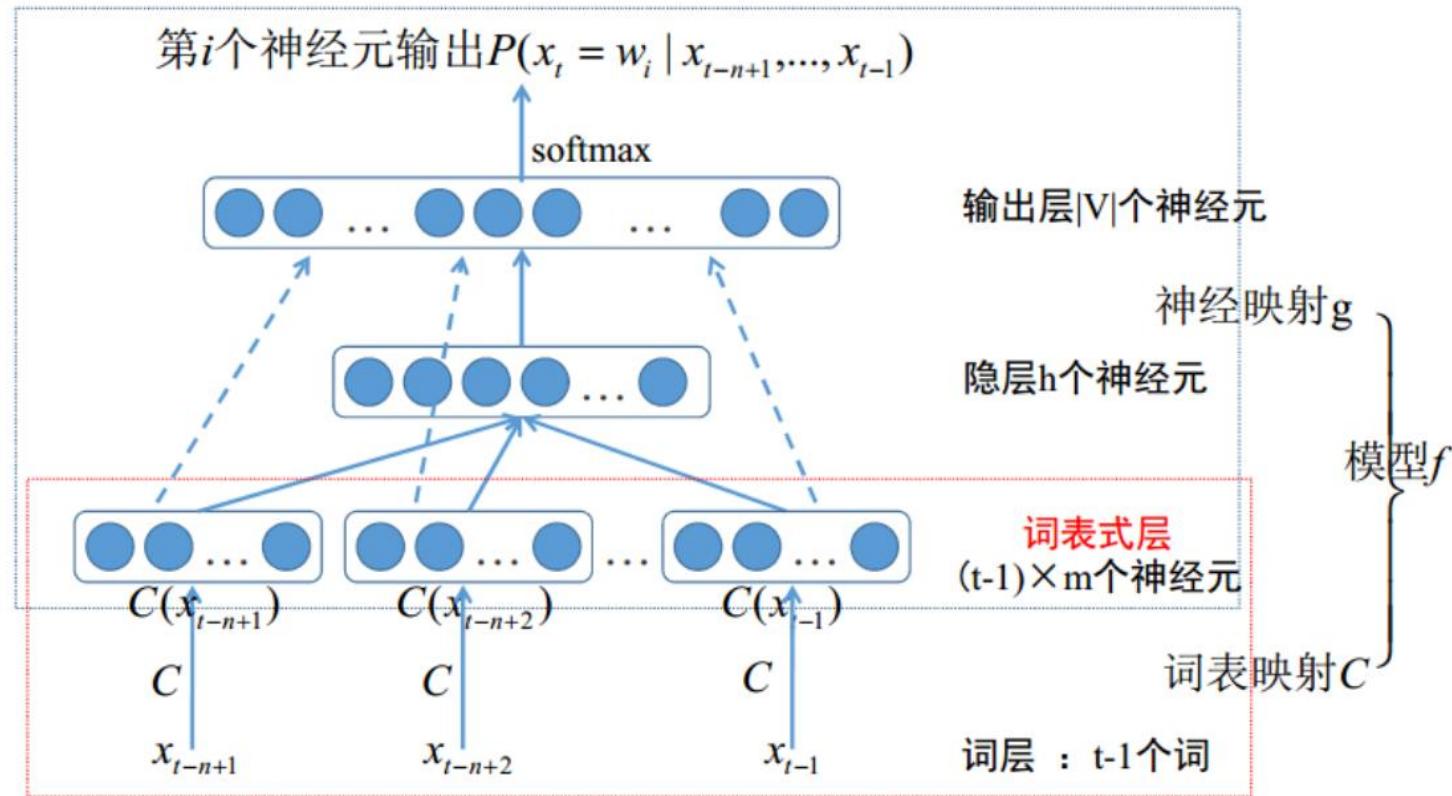
■基于神经网络的词向量学习

■基于大规模真实数据：LM的副产品

- [Bengio2003] Yoshua Bengio, Rejean Ducharme, Pascal Vincent,Christian Jauvin, Jaz K, Thomas Hofmann, Tomaso Poggio, and John Shawe-taylor. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155. 2003.
- [Mnih2008]Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 641–648, New York, NY, USA. ACM. Mnih, A., & Hinton, G. (2008). A Scalable Hierarchical Distributed Language Model. *NIPS2008* (pp. 1–8).
- [Turian2010]Turian, J., & Ratinov, L. (2010). Word representations : A simple and general method for semi-supervised learning. *ACL2010* (pp. 384–394)
-

■ 神经概率语言模型[Bengio2003] :

■ 在获得语言模型(后续)的同时获得了词的分布式表示



- 基于神经网络的词分布表示获取技术：直接面向词向量学习
 - 模型复杂度更小、数据规模更大下学习词表示
 - [Mikolov2013ICLRworkshop]Tomas Mikolov, Greg Corrado, Kai Chen, Jeffrey Dean, Efficient Estimation of Word Representation in Vector Space, ICLR2013 workshop
 - [Mikolov2013NIPS]Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems. 2013
 - [Mikolov2013NAACL]Tomas Mikolov, Wen-tau Yih, Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. Proceedings of NAACL-HLT 2013, pages 746–751
 - <https://code.google.com/p/word2vec/>
 -

■ 2个模型

■ CBOW

■ Skip-gram

■ 2种训练方法

■ 层次softmax

■ 负采样

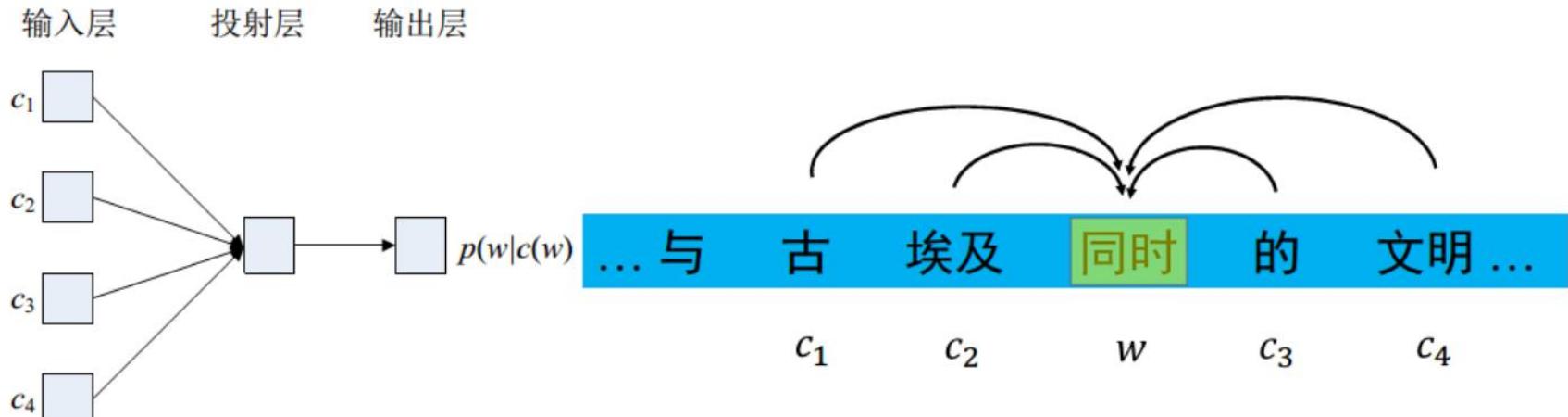
■ 基于向量的方法

- 基于统计的高维向量及其降维
- 基于预测的低维向量
 - CBOW + Hierarchical SoftMax
 - Skip-Gram + Negative Sampling
- 词向量评估
- 问题与发展
- 词向量应用

■CBOW(Continuous Bag-Of-Words)

■出发点：通过上下文的环境来预测当前词

■对于目标词 w 及其上下文 $C(w) = (c_1, c_2, c_3, c_4)$ ： $p(w|C(w))$

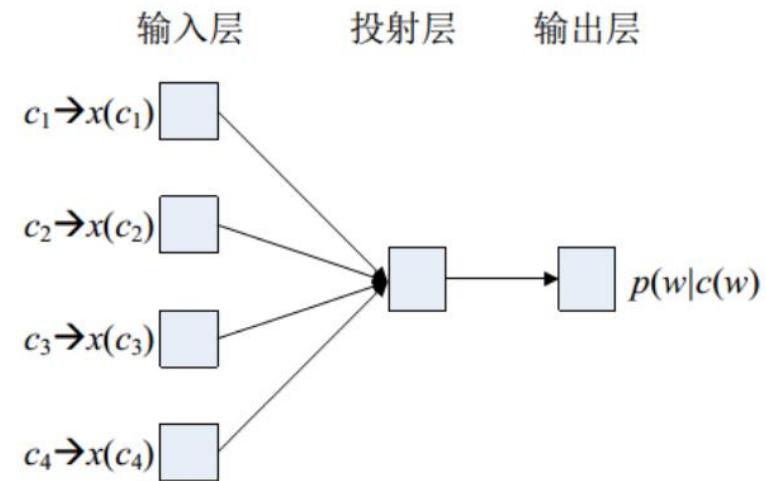
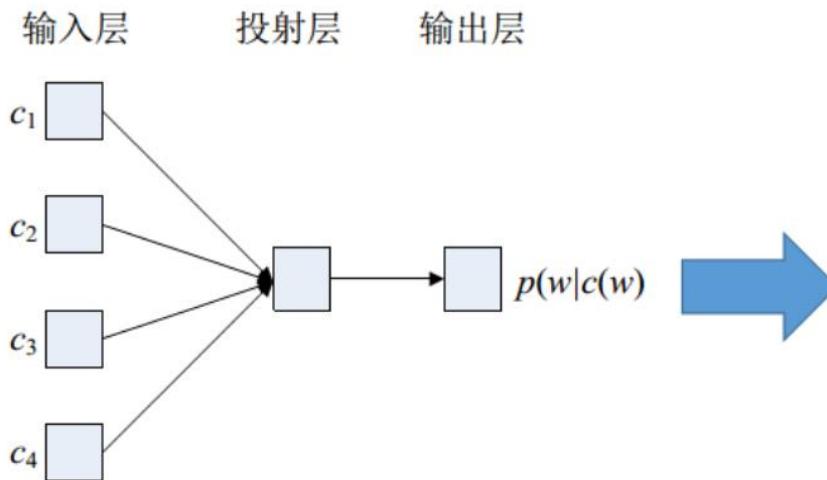


- 1. 此处以词向量学习为例，其他单元(子词)的向量表示学习类似；
- 2. 此处以及后文中上下文以目标词前后2词(上下文窗口k=2)为例；

■CBOW(Continuous Bag-Of-Words)

■出发点：通过上下文的环境来预测当前词

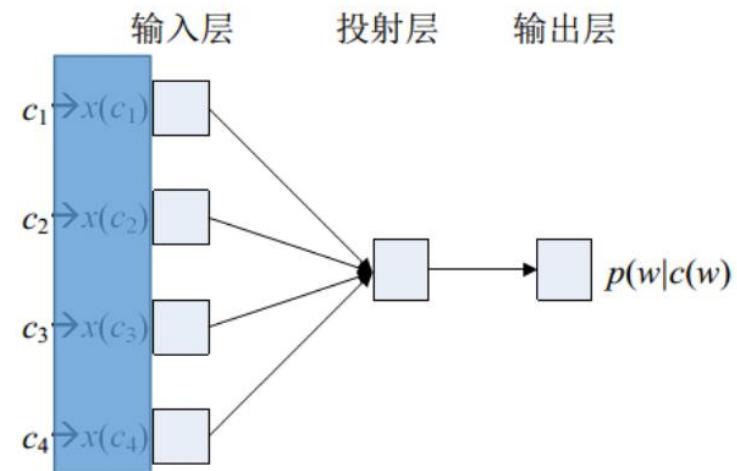
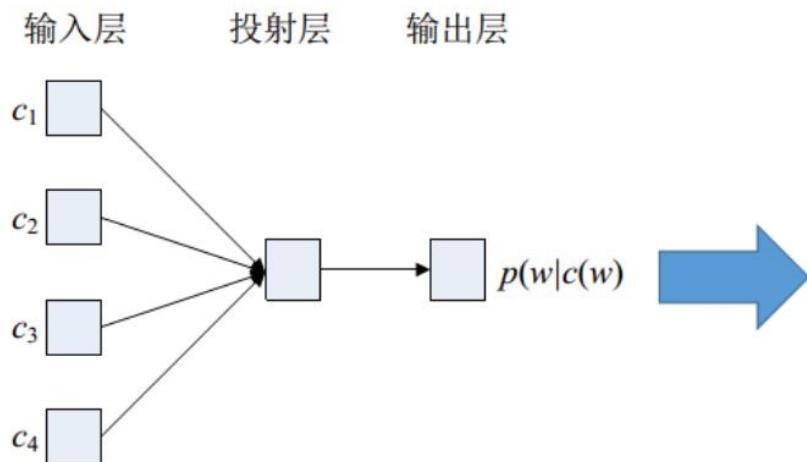
■对于目标词 w 及其上下文 $C(w) = (c_1, c_2, c_3, c_4)$ ： $p(w | C(w))$



■CBOW(Continuous Bag-Of-Words)

■出发点：通过上下文的环境来预测当前词

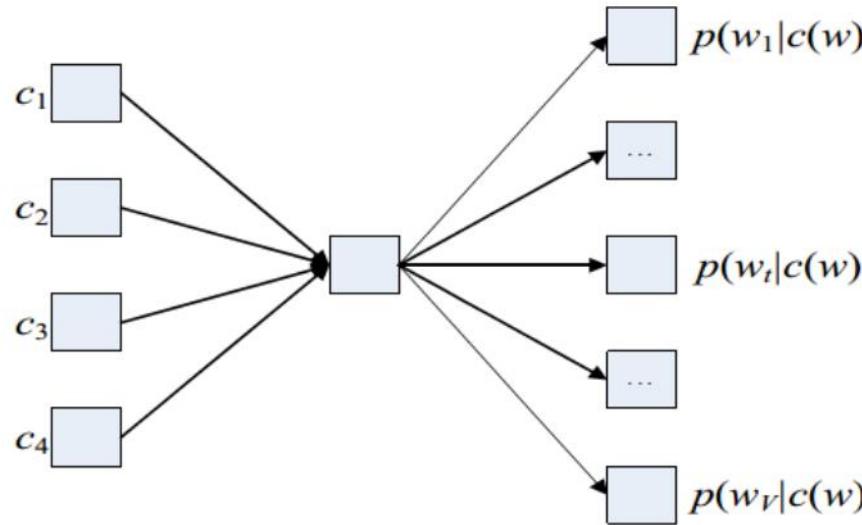
■对于目标词 w 及其上下文 $C(w) = (c_1, c_2, c_3, c_4)$ ： $p(w|C(w))$



后续其他模型如不特别说明时输入都是向量

- 理想目标: $p(w|C(w))$ 极大 (至少 $P(w|C(w)) > P(\text{非}w|C(w))$)
- 但是 非 w 是什么?
- 一个直接的方案是构建V类的softmax, 所有其他词均为非 w :

输入层 投射层 输出层: V个输出



- 但是模型复杂度高
- 如何更高效?

■ 层次softmax

输入层

c_1

c_2

c_3

c_4

投射层

T_1

输出层：哈夫曼树

$p(w_1|c(w))$

$p(w_2|c(w))$

$p(w|c(w))$

$p(w_3|c(w))$

将一个V分类分解为一系列的二分到达某个词

■问题：

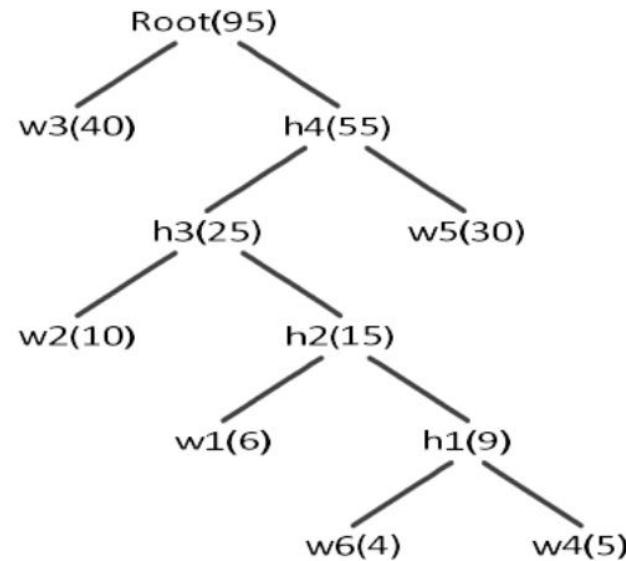
- 一、二分树有很多种，如何构造合适的二分树？
- 二、分到叶子节点的 w 要经过多次二分，如何设计二分类器，并计算最终分到叶子节点 w 的概率？
- 三、如何训练所有二分类器的参数？

■一、二分树：基于词频的哈夫曼树

■基于语料中的词频构建

■ $w_1: f(w_1), w_2: f(w_2), \dots, w_n: f(w_n)$

■例：如有 $f(w_1)=6, f(w_2)=10, f(w_3)=40, f(w_4)=5, f(w_5)=30, f(w_6)=4$ ，
则其哈夫曼树如下：

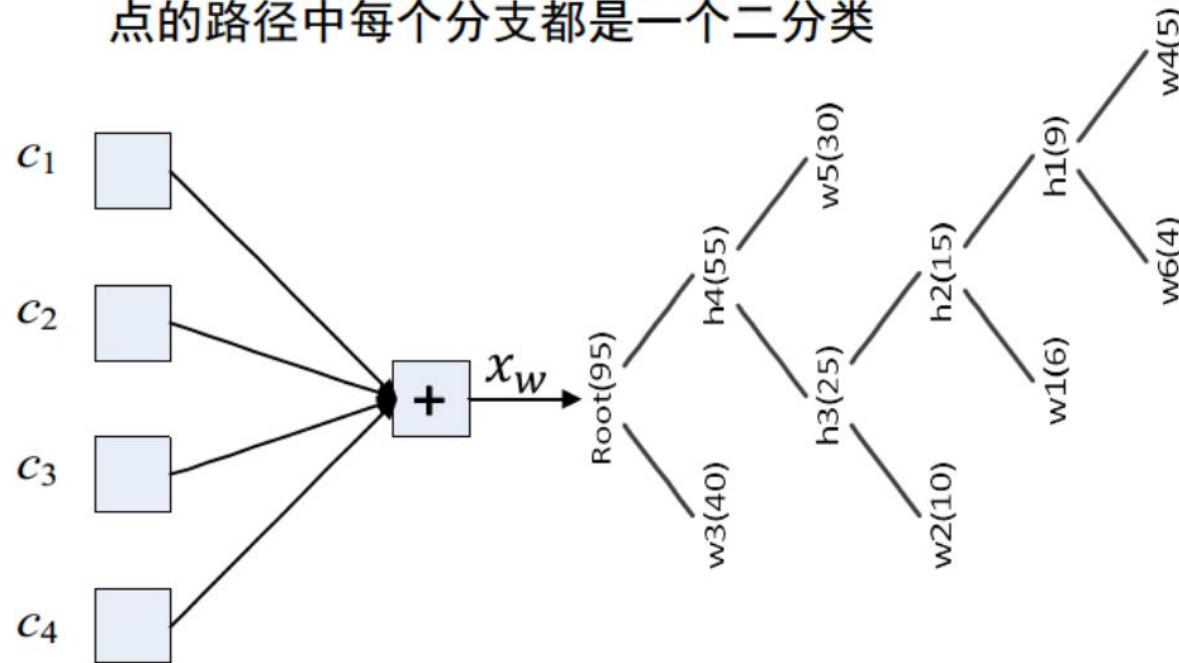


从根节点出发，频率高的
词具有较短的路径

■二、层次softmax算法

■投射层：简化为各个输入向量相加

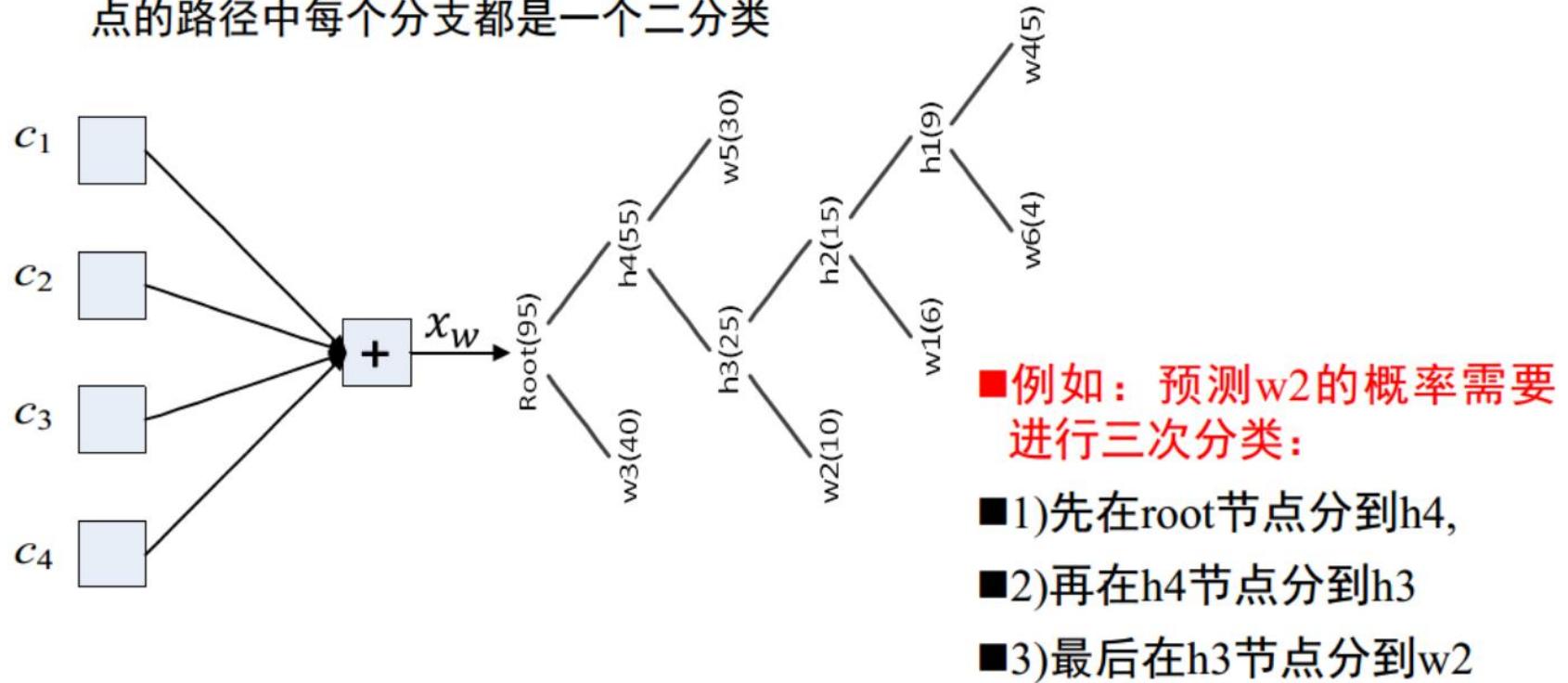
■输出层：一个哈夫曼树，树的叶节点是分到某个词的概率，从根节点到叶节点的路径中每个分支都是一个二分类



■二、层次softmax算法

■投射层：简化为各个输入向量相加

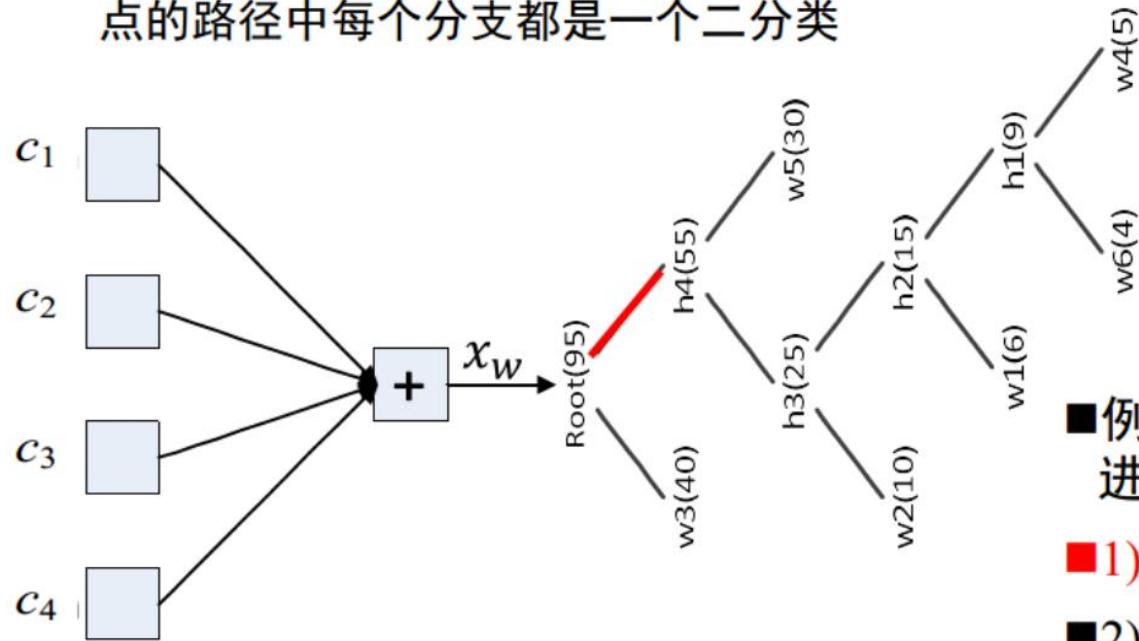
■输出层：一个哈夫曼树，树的叶节点是分到某个词的概率，从根节点到叶节点的路径中每个分支都是一个二分类



■二、层次softmax算法

■投射层：简化为各个输入向量相加

■输出层：一个哈夫曼树，树的叶节点是分到某个词的概率，从根节点到叶节点的路径中每个分支都是一个二分类



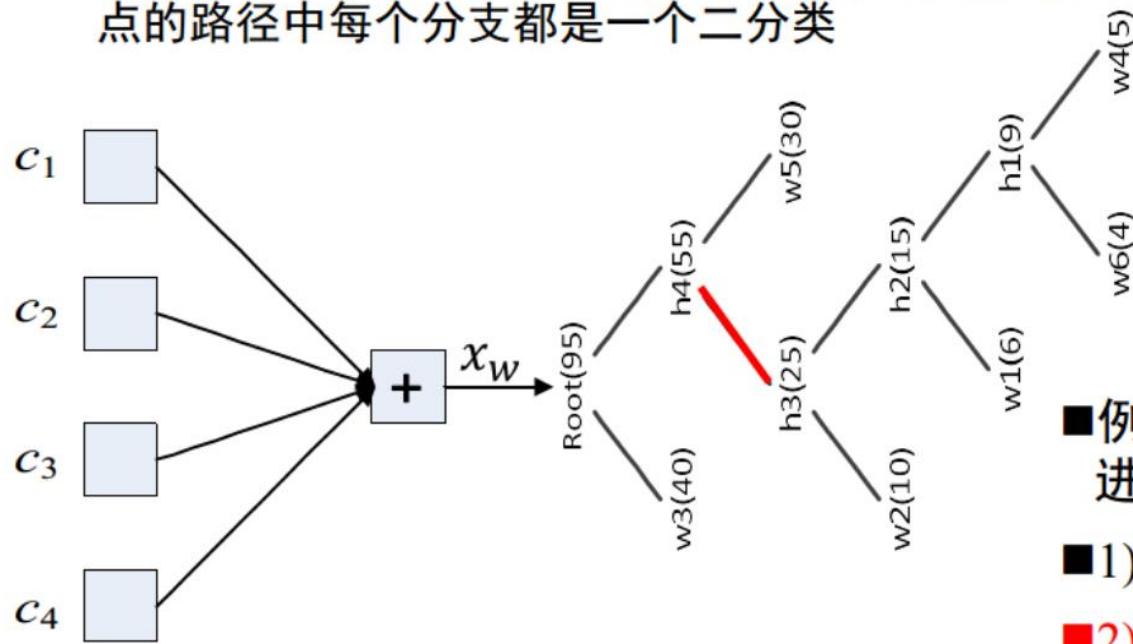
■例如：预测 w_2 的概率需要进行三次分类：

- 1)先在root节点分到 h_4 ,
- 2)再在 h_4 节点分到 h_3
- 3)最后在 h_3 节点分到 w_2

■二、层次softmax算法

■投射层：简化为各个输入向量相加

■输出层：一个哈夫曼树，树的叶节点是分到某个词的概率，从根节点到叶节点的路径中每个分支都是一个二分类



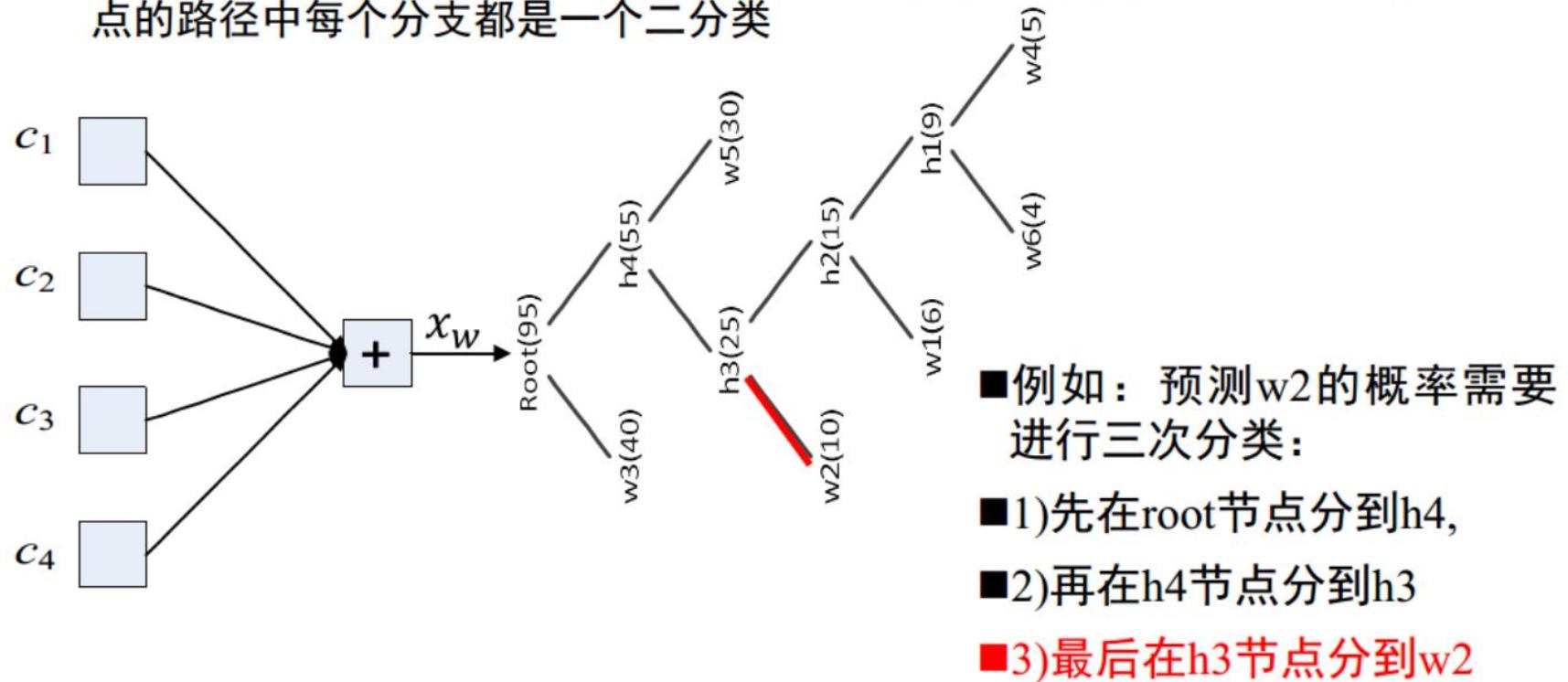
■例如：预测w2的概率需要进行三次分类：

- 1)先在root节点分到h4,
- 2)再在h4节点分到h3
- 3)最后在h3节点分到w2

■二、层次softmax算法

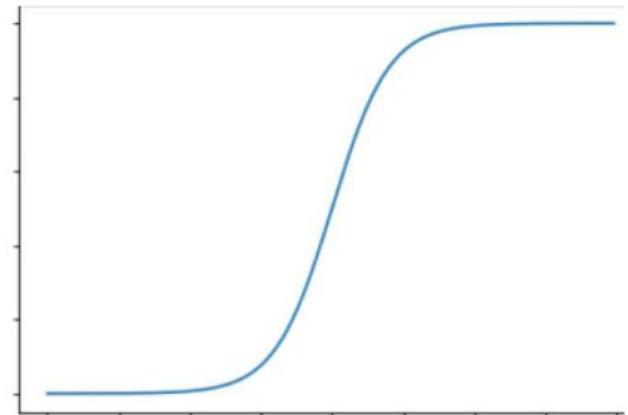
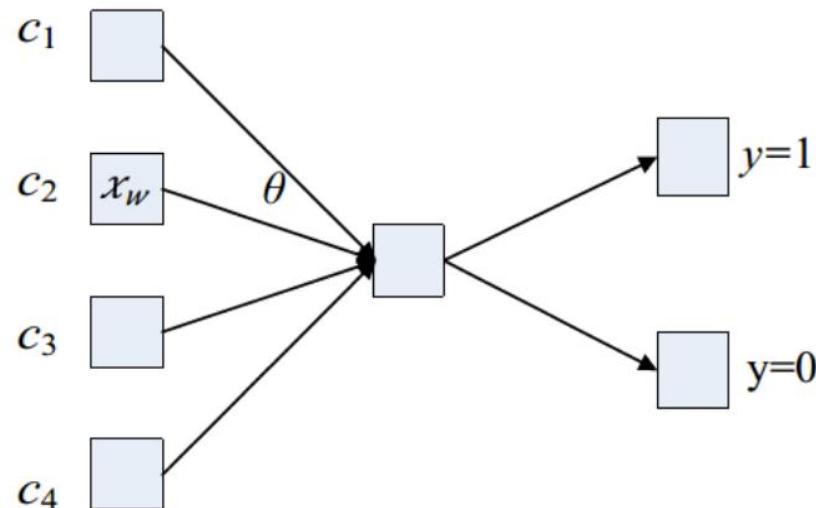
■投射层：简化为各个输入向量相加

■输出层：一个哈夫曼树，树的叶节点是分到某个词的概率，从根节点到叶节点的路径中每个分支都是一个二分类

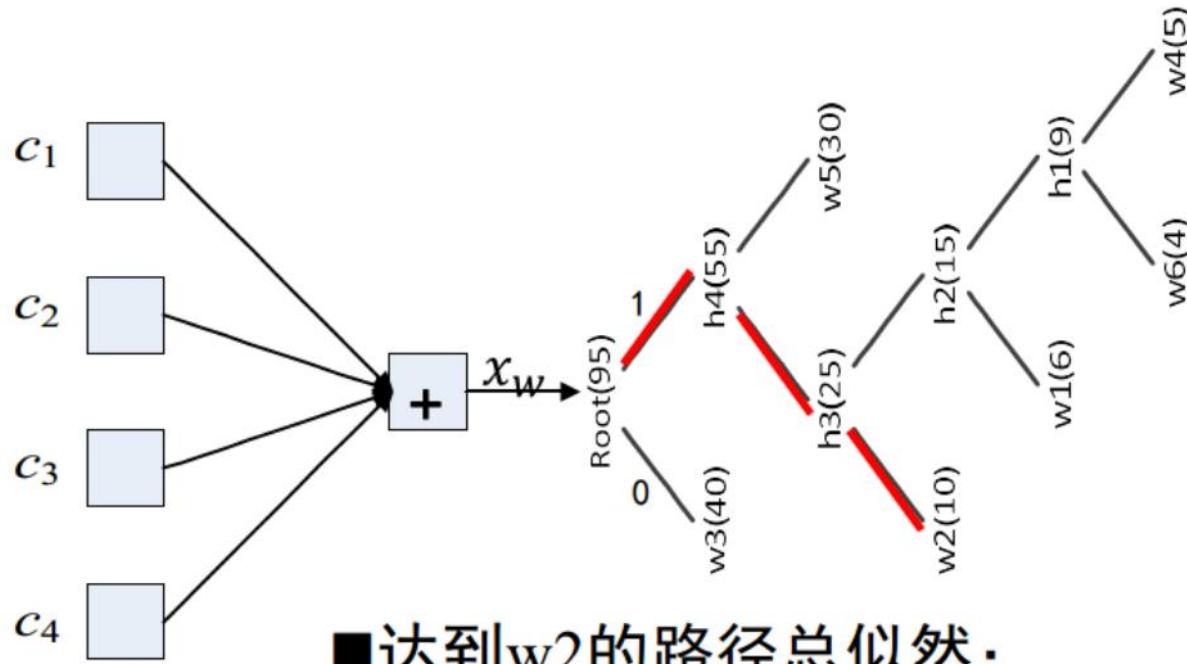


■ 每个二分类($y=\{0,1\}$)都采用logistic回归：

$$\blacksquare P(y|x_w, \theta) = [\sigma(x_w \cdot \theta)]^{1-y} \cdot [1 - \sigma(x_w \cdot \theta)]^y$$



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



■ 达到w2的路径总似然：

$$\begin{aligned}
 & [\sigma(x_w \cdot \theta_{root})]^{1-1} \cdot [1 - \sigma(x_w \cdot \theta_{root})]^1 \\
 & * [\sigma(x_w \cdot \theta_{h4})]^{1-0} \cdot [1 - \sigma(x_w \cdot \theta_{h4})]^0 \\
 & * [\sigma(x_w \cdot \theta_{h3})]^{1-0} \cdot [1 - \sigma(x_w \cdot \theta_{h3})]^0
 \end{aligned}$$

■ 目标：极大化上述似然

■一般地，每个二分类的概率：

$$\begin{aligned} \blacksquare P(d_j^w | x_w, \theta_{j-1}^w) = \\ [\sigma(x_w^T \theta_{j-1}^w)]^{1-d_j^w} \cdot [1 - \sigma(x_w^T \theta_{j-1}^w)]^{d_j^w} \end{aligned}$$

■ x_w 为输入词向量

■ θ_{j-1}^w 表示从j-1层到j层的参数

■ d_j^w 表示第j层二分类结果，对于哈夫曼树， d_j^w 的值取1或0。

■ 训练的目标就是极大化到达正确词的路径的概率

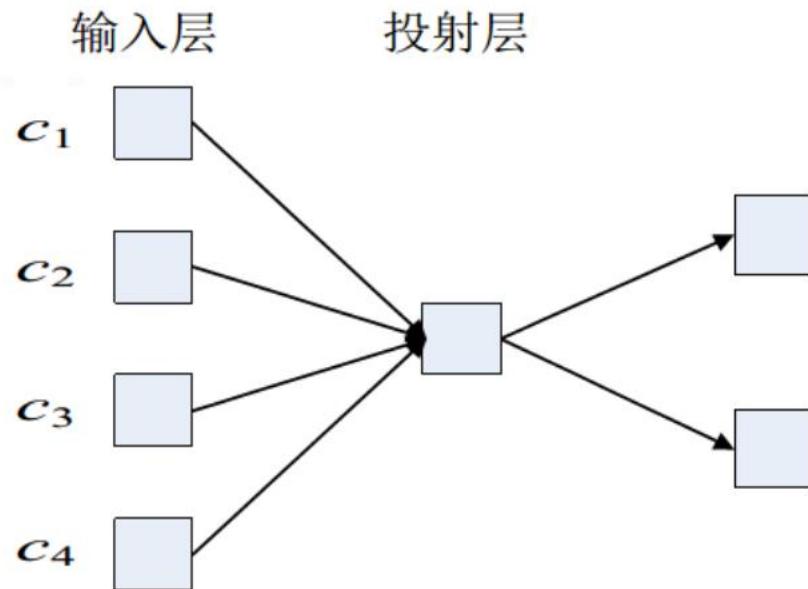
$$\blacksquare P(w|C(w)) = \prod_{j=2}^{l^w} P(d_j^w | x_w, \theta_{j-1}^w)$$

$$\blacksquare P(d_j^w | x_w, \theta_{j-1}^w) = [\sigma(x_w^T \theta_{j-1}^w)]^{1-d_j^w} \cdot [1 - \sigma(x_w^T \theta_{j-1}^w)]^{d_j^w}$$

■ 要使这个概率最大化可以采用随机梯度下降法，对参数 x_w 和 θ_{j-1}^w 进行更新

■ 注意 x_w 为多个词的向量之和，直接把 x_w 的更新梯度用在各个词向量上。

- 层次softmax方法问题：
 - 树的结构影响大
 - 树的训练复杂性高
- 能否更简洁：直接解决用非w而不是用所有w？

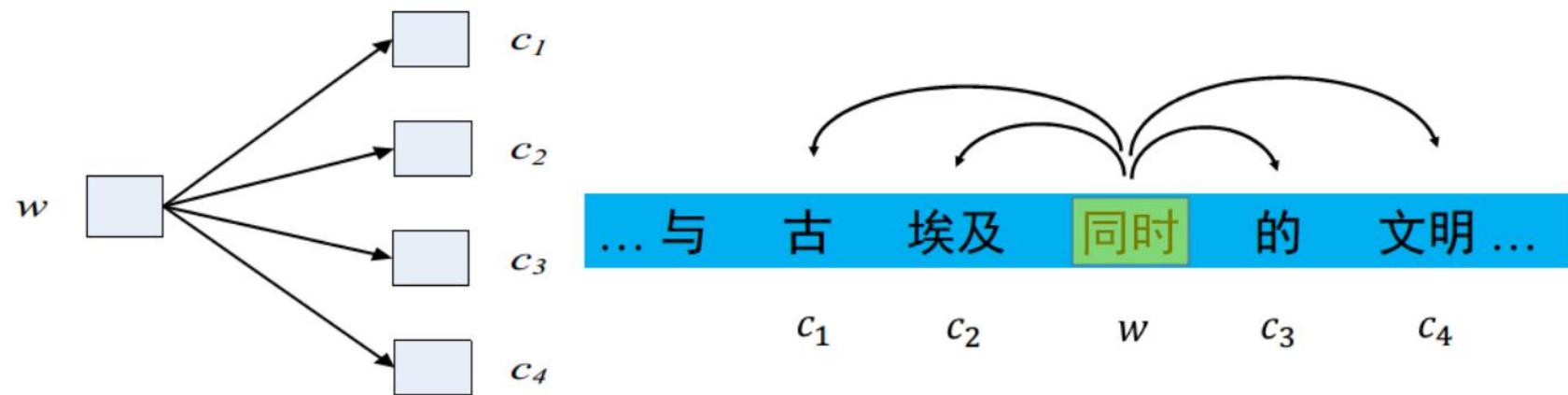


■ 基于向量的方法

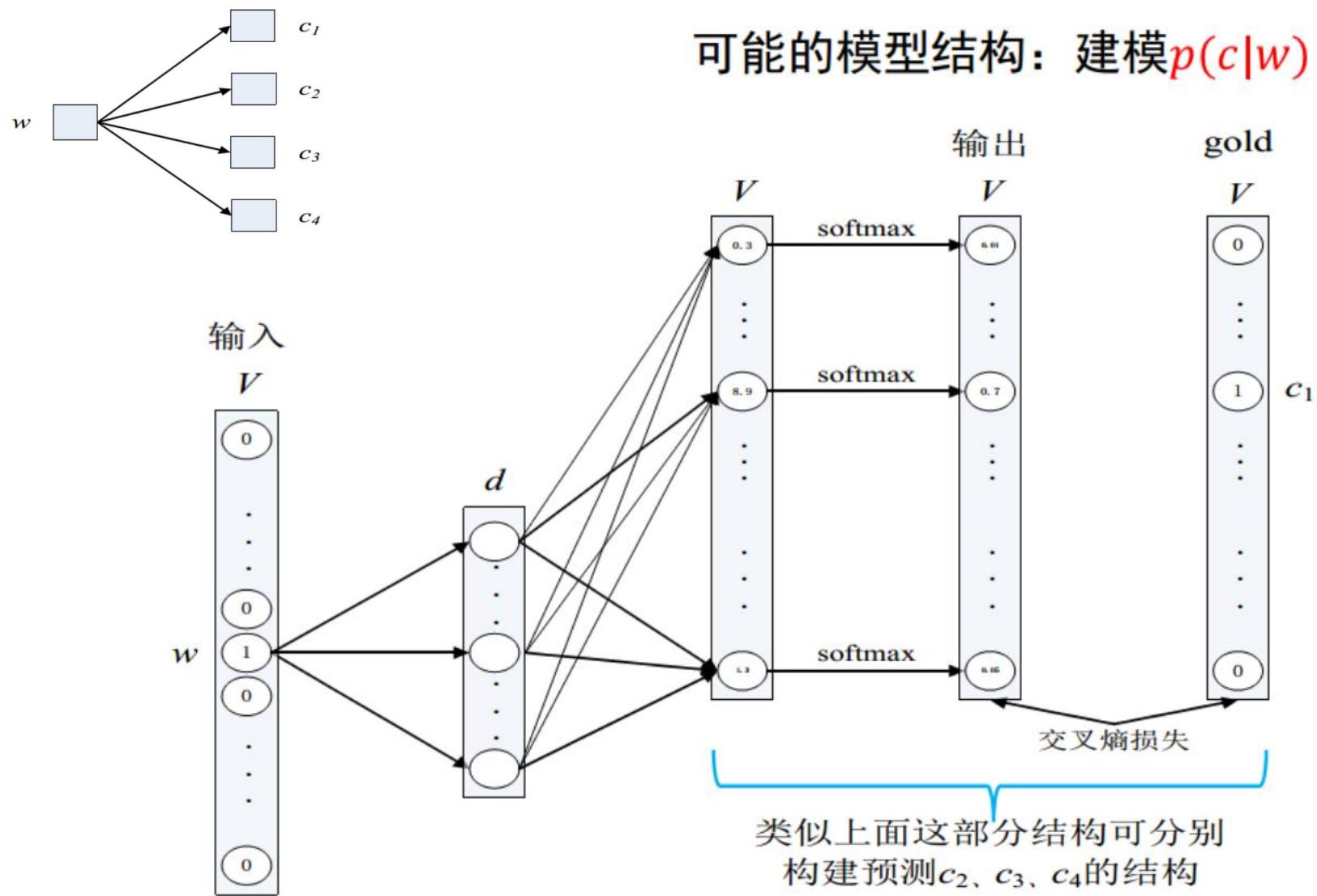
- 基于统计的高维向量及其降维
- 基于预测的低维向量
 - CBOW + Hierarchical SoftMax
 - Skip-Gram + Negative Sampling
- 词向量评估
- 问题与发展
- 词向量应用

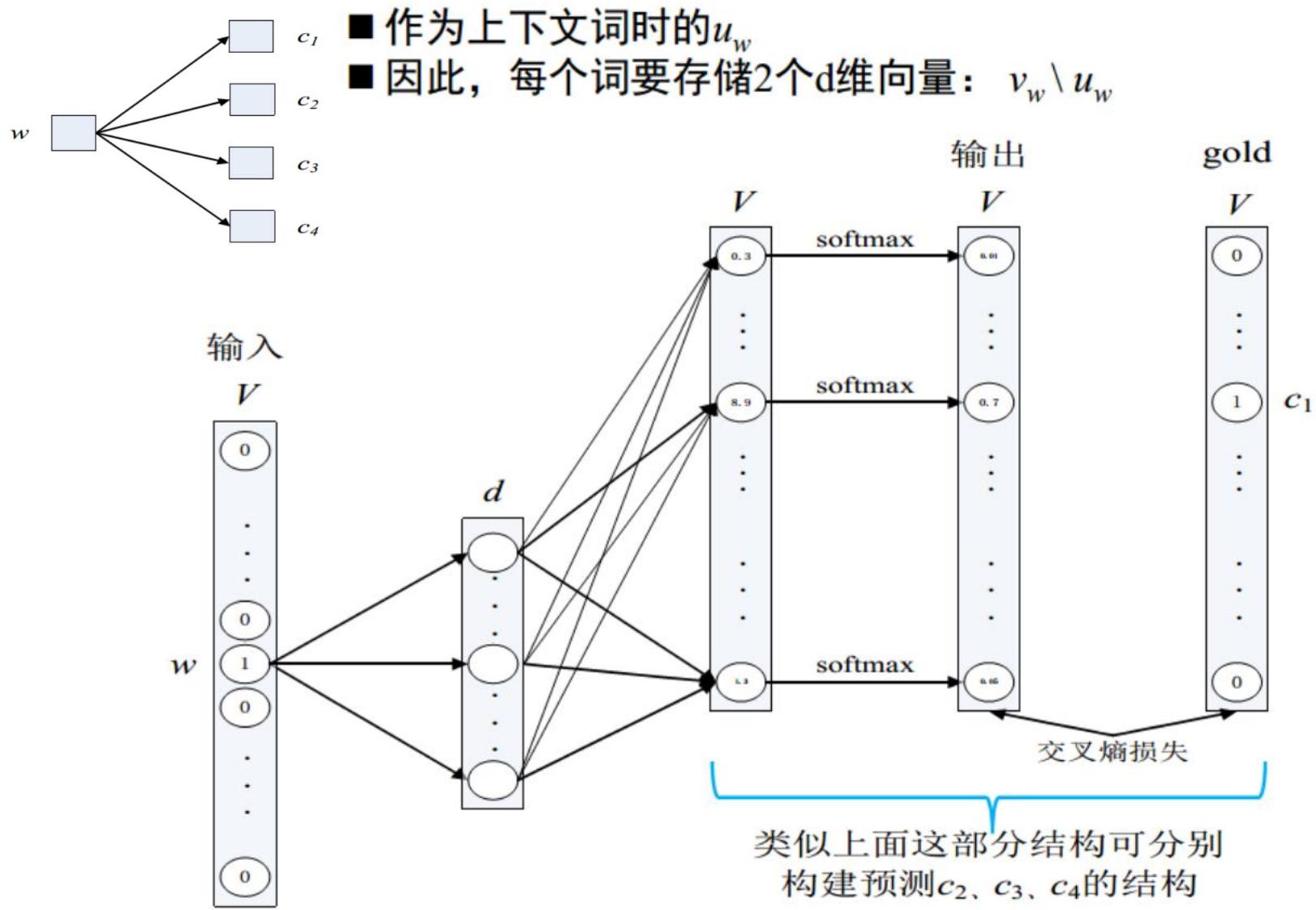
■ Skip-gram模型：预测当前词的上下文词

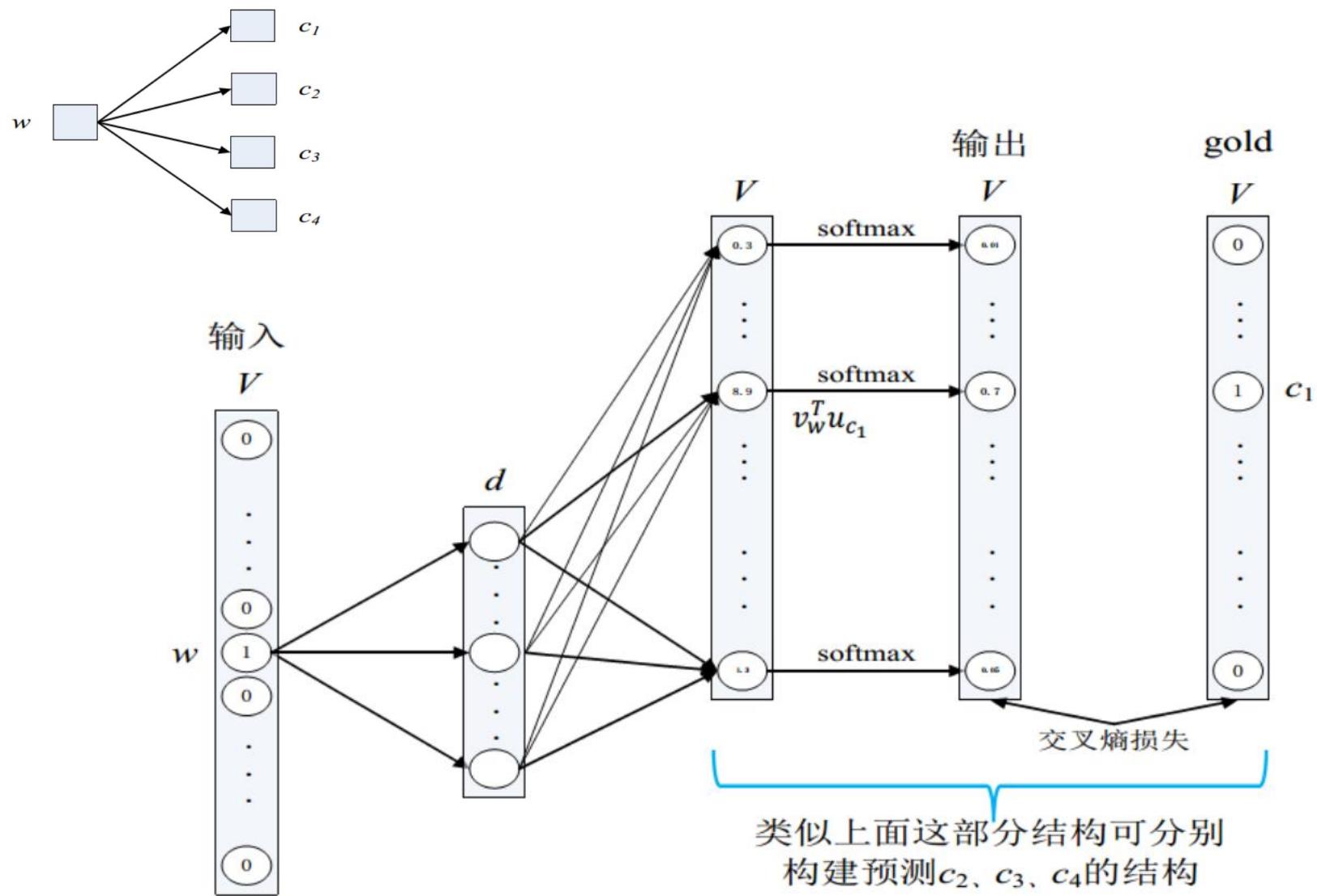
■ 将词向量学习融合进去

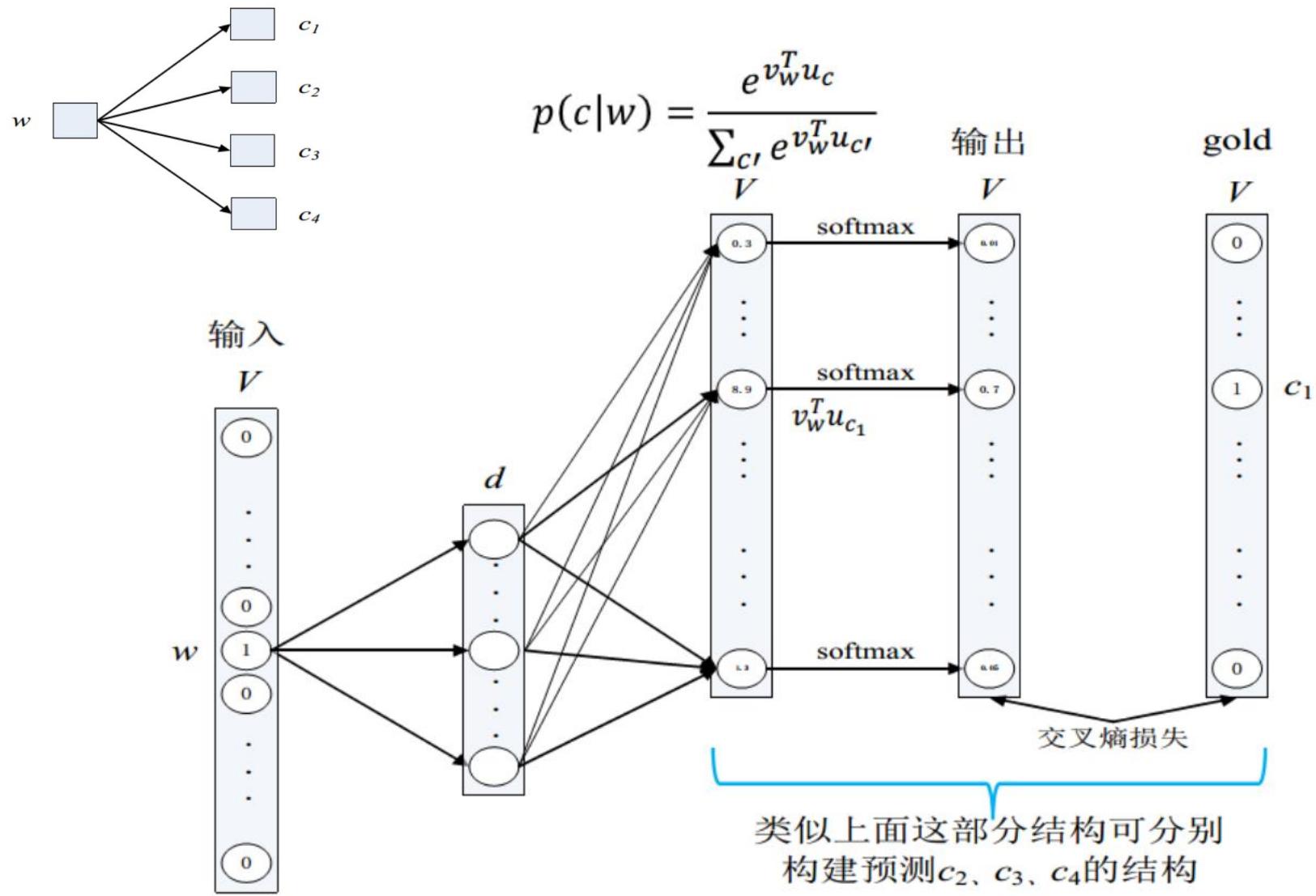


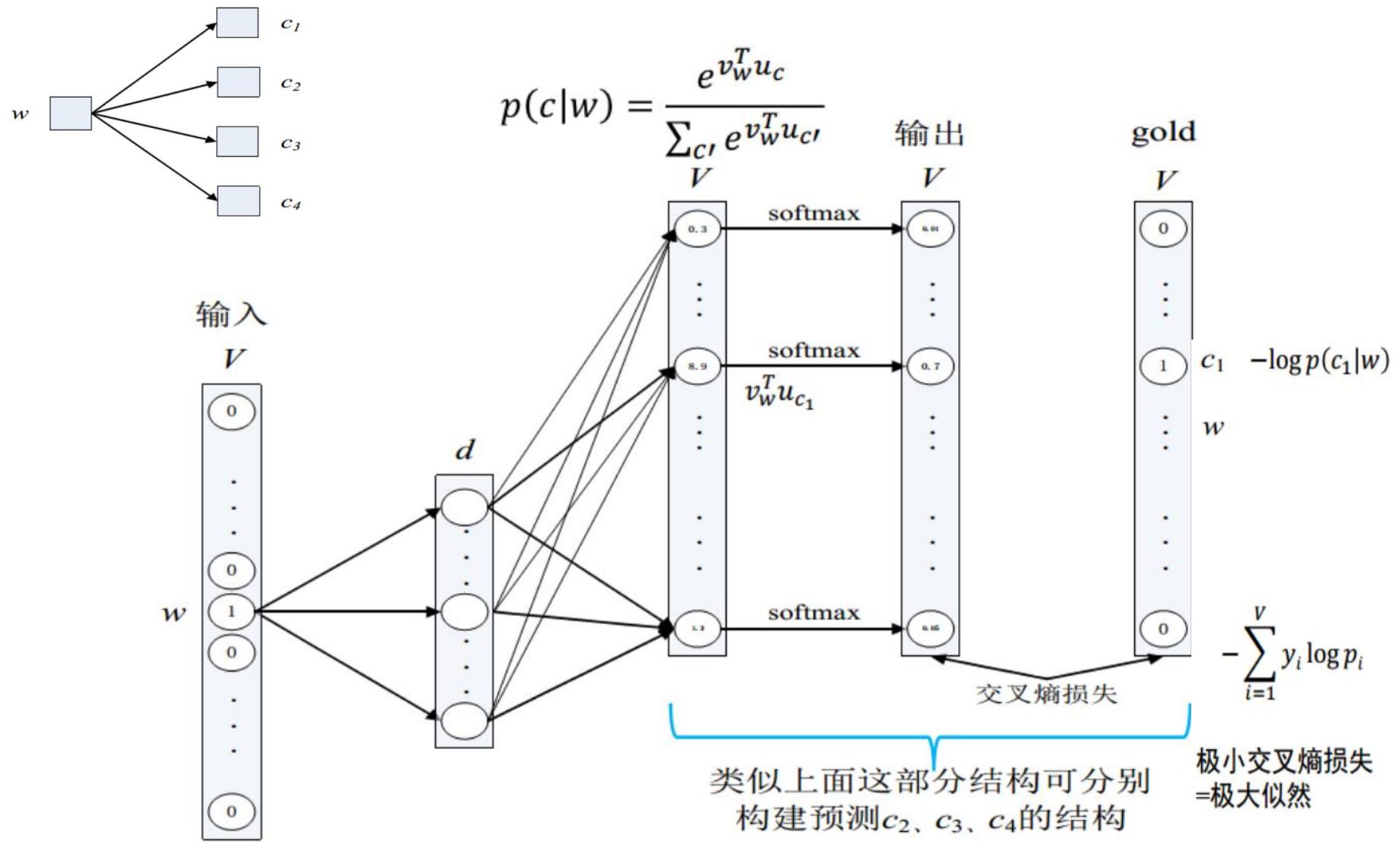
可能的模型结构：建模 $p(c|w)$











■ 给定单个样本 $(w; c_1, c_2, c_3, c_4)$, 仅以 $(w; c_1)$ 部分为例:

■ 前向:

■ 1. 计算 $v_w u_c$, c 遍历 $\{w_1, \dots, w_V\}$, 其中有某 $w_i = c_1$

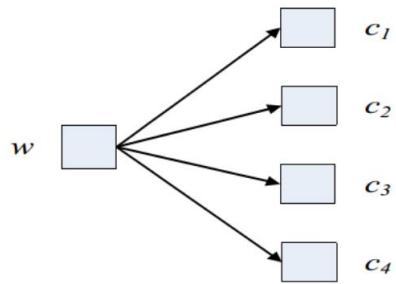
■ 2. 为每个 c 计算 $p(c|w) = \frac{e^{v_w^T u_c}}{\sum_{c'} e^{v_w^T u_{c'}}}$, $c' = \{w_1, \dots, w_V\}$

■ 3、计算损失函数 $J = -\log p(c_1|w) = -\log \frac{e^{v_w^T u_{c_1}}}{\sum_{c'} e^{v_w^T u_{c'}}}$

■ 反向:

■ 1. $\frac{\partial J}{\partial v_w}, \frac{\partial J}{\partial u_c}$ (c 是某一个具体的 $c' = \{w_1, \dots, w_V\}$)

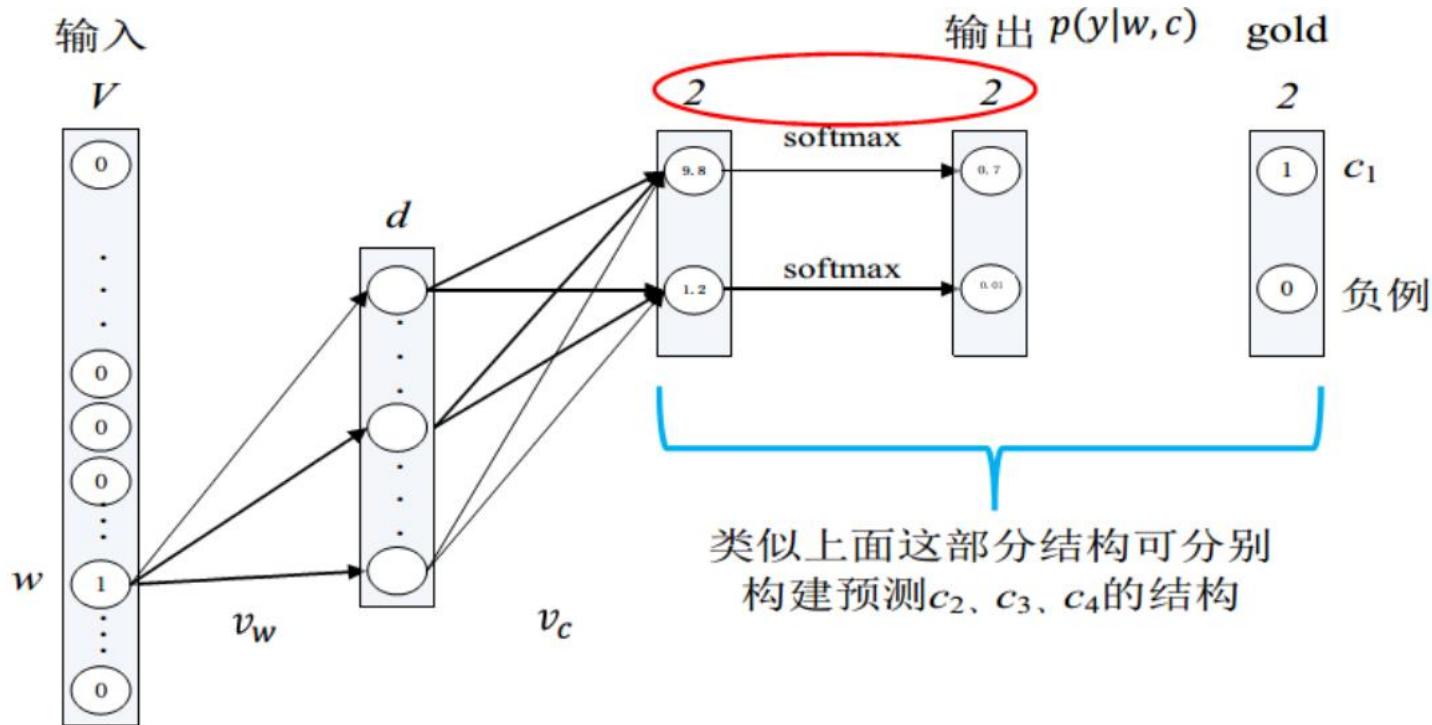
■ 2. $v_w \leftarrow v_w - \eta \frac{\partial J}{\partial v_w}, u_c \leftarrow u_c - \eta \frac{\partial J}{\partial u_c}$ (η 为学习率)

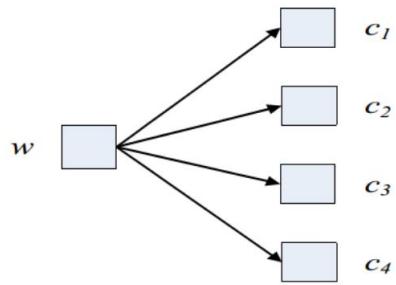


转换为二分类问题(NCE: Noise Contrastive Estimation):

1. $p(c|w) \rightarrow p(y|w, c)$: $p(y|w, c) = \begin{cases} 1 & w, c \text{ 为正例} \\ 0 & w, c \text{ 为负例} \end{cases}$

2. 负样例来自一个噪声分布；
3. 一个正例抽取k个负例。





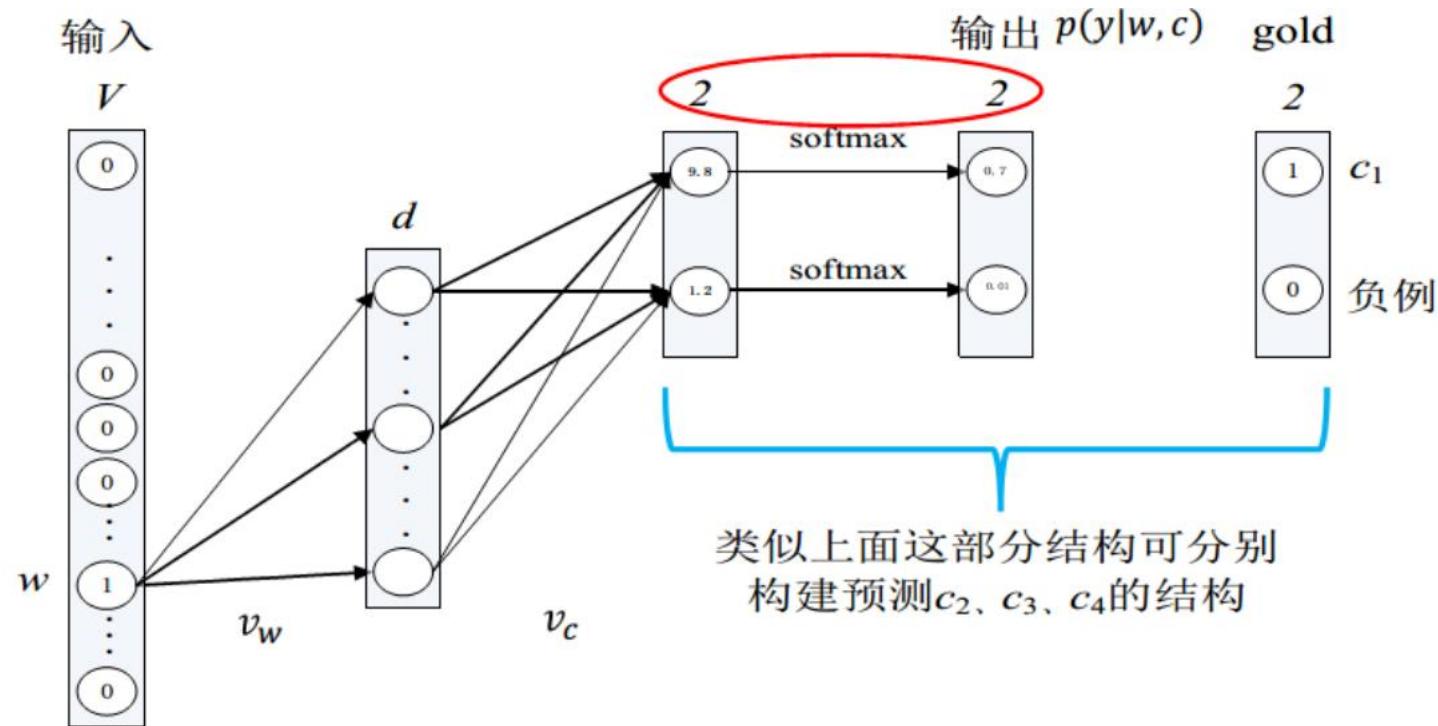
几点简化(NS)

1、二分类采用logistic回归：

$$p(y = 1|w, c) = \sigma(v_w^T u_c) \quad p(y = 0|w, c) = \sigma(-v_w^T u_c)$$

2、负例采样分布： $p_\alpha(\tilde{c}) = \frac{\text{count}(\tilde{c})^\alpha}{\sum_{\tilde{c}'} \text{count}(\tilde{c}')^\alpha}$ $\alpha = 3/4$

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

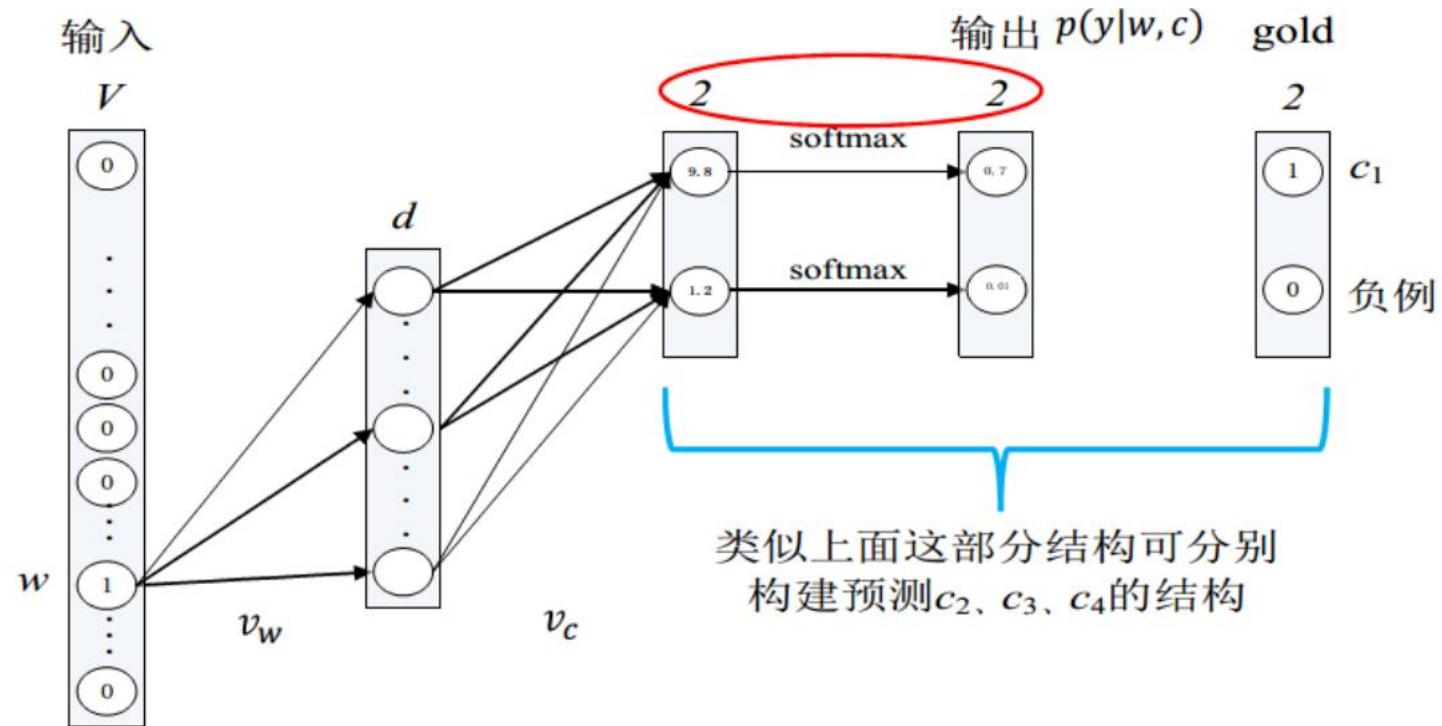
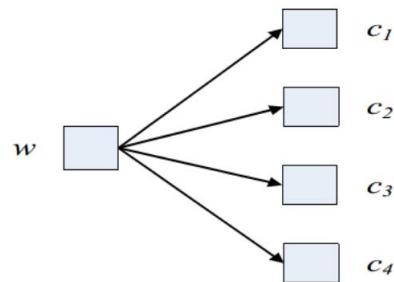


几点简化(NS)

3、一个正例配k个负例(数据小5-20， 数据多2-5)

Negative Sampling损失函数：

$$J = \log\sigma(v_w^T u_c) + \sum_{i=1}^k \log\sigma(-v_w^T u_{\tilde{c}_i})]$$



■误差反传 + 梯度下降

- $\frac{\partial J}{\partial v_w}, \frac{\partial J}{\partial u_c}, \frac{\partial J}{\partial u_{\tilde{c}_i}};$

- $v_w \leftarrow v_w - \eta \frac{\partial J}{\partial v_w}$

- $u_c \leftarrow u_c - \eta \frac{\partial J}{\partial u_c}$

- $u_{\tilde{c}_i}' \leftarrow u_{\tilde{c}_i} - \eta \frac{\partial J}{\partial u_{\tilde{c}_i}}$

■上述为一个样本

■多个样本累积后

- $J = \sum_{w=1}^{batch_size} (\log \sigma(v_w^T u_c) + \sum_{i=1}^k \log \sigma(-v_w^T u_{\tilde{c}_i}))$

■训练样本构建示例：上下文窗口=2，负例数=2

语料：... 与 古 埃及 同时 的 文明 ...

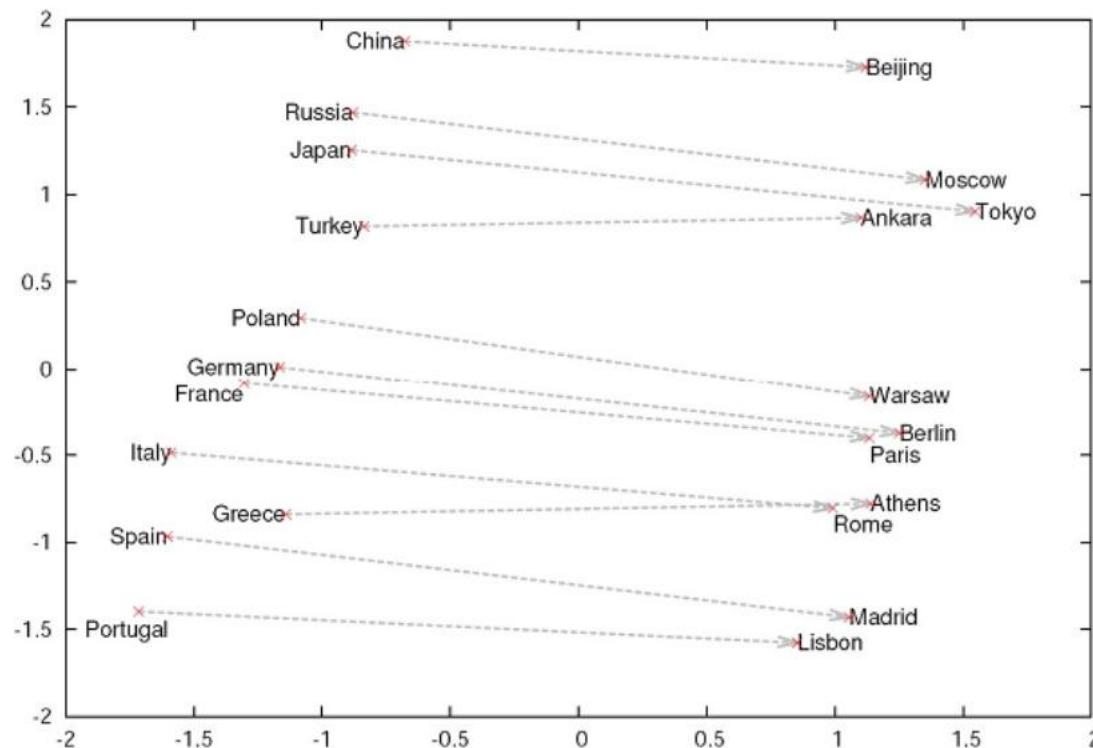
输入:w	正例:c	负例: \tilde{c}	
同时	古	非“古”： 基于 $p_\alpha(\tilde{c}) = \frac{\text{count}(\tilde{c})^\alpha}{\sum_{\tilde{c}'} \text{count}(\tilde{c}')^\alpha}$ 采样2个词	一个样本
	埃及	非“埃及”： 基于 $p_\alpha(\tilde{c}) = \frac{\text{count}(\tilde{c})^\alpha}{\sum_{\tilde{c}'} \text{count}(\tilde{c}')^\alpha}$ 采样2个词	一个样本
	的	非“的”： 基于 $p_\alpha(\tilde{c}) = \frac{\text{count}(\tilde{c})^\alpha}{\sum_{\tilde{c}'} \text{count}(\tilde{c}')^\alpha}$ 采样2个词	一个样本
	文明	非“文明”： 基于 $p_\alpha(\tilde{c}) = \frac{\text{count}(\tilde{c})^\alpha}{\sum_{\tilde{c}'} \text{count}(\tilde{c}')^\alpha}$ 采样2个词	一个样本

- 采用下采样(subsampling)技术降低无效语料规模
 - 在训练模型时，高频词频繁作为 w 、 c ，尤其像“的、地、得\the、a”等等对于形成其他词的表示作用不是很大的高频词
 - 采用下采样技术以减少这些词作为样本的机会
 - 遇到这些词时以某个概率 $p(w)$ 保持， $1-p(w)$ 扔掉。
 - 例如：
$$p(w) = \left(\sqrt{\frac{f(w)}{0.001}} + 1 \right) \frac{0.001}{f(w)}$$
 - $f(w)$ 为该词在所用训练语料中的频率

- 向量选择：
 - 每个词 w 有两个向量
 - 作为目标词 vs 作为上下文词
 - 几种可能的选择：
 - 其作为目标词的向量
 - 其作为目标词的向量加上其作为上下文词的向量
 - 其作为目标词的向量拼接其作为上下文词的向量，形成2d维的词向量
- 算法的计算复杂性
 - 正比与语料规模 $|C|$ ， $\mathcal{O}(|C|)$ 。

■ 分布式词表示：可视化

■ Mikolov 2013：一些1000维skip-gram向量通过PCA降到2维的可视化：国家和首都



■ 基于向量的方法

- 基于统计的高维向量及其降维
- 基于预测的低维向量
 - CBOW + Hierarchical SoftMax
 - Skip-Gram + Negative Sampling
- 词向量评估
- 问题与发展
- 词向量应用

■词表示评估

■外部任务：用于下游任务

- 文本分类\语言理解\机器翻译\.....

■词表示评估

- 外部任务：用于下游任务

- 内在任务：

- 词相似度任务

- 词类比任务

■词相似度任务

■数据集：

■汉语：PKU500、CWE297,CWE240

■英语：WordSim353、RW、MEN、

■<http://alfonseca.org/eng/research/wordsim353.html>

■德语：ZG222、Gur250

词1	词2	人工打分(10人打分的均值, [1,10])
love	sex	6.77
tiger	cat	7.35
tiger	tiger	10.00
drink	car	3.04
drink	ear	1.31
drink	mouth	5.96
smart	student	4.62
glass	magician	2.08

词1	词2	人工打分
没戏	没辙	4.9
物理	质子	2.6
GDP	生产力	6.5
结盟	无理取闹	1.1
由此	通过	3.4
商业	工业	4.8
忌妒	妒嫉	9.2
随意	随便	7.4

■词类比任务：给出三个找第四个

■句法类比(syntactic analogy)

■big : bigger = small : smaller

■语义类比(semantic analogy)

■France : Paris = Germany : Berlin

■数据集：

■[Mikolov2013NAACL] 创建了句法类比数据集，含8000条。

■semantic analogy数据采用的是SemEval-2012 Task 2, Measuring Relation Similarity，由[Jurgens2012SemEval]创建。

■问题与发展

■每次用局部上下文进行预测，全局统计信息？

■问题：不是基于对于整个语料的同现统计数据，只是扫描局部窗口中的词，因此不能利用整体数据的统计特征。From [Pennington2014EMNLP]

- 全局统计信息：
 - 词(目标词)-词(上下文)同现概率矩阵
- 和直接基于同现概率矩阵分解方法的比较：
 - Omer Levy, Yoav Goldberg. Neural Word Embedding as Implicit Matrix Factorization. NIPS2014
 - Benjamin Wilson. Skip-gram isn't Matrix Factorisation. 2016
 - Ryan Cotterell, Adam Poliak, Benjamin Van Durme, and Jason Eisner. Explaining and Generalizing Skip-Gram through Exponential Family Principal Component Analysis. EACL2017.
 -
- Pennington等[Pennington2014EMNLP]认为比同现概率更有效的是同现概率的比，基于此提出了GloVe(Global Vectors)

■ GloVe动机：同现概率 \rightarrow 同现概率比

■ 例子：要区分ice和steam

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

■ 表中ratio来自对6 billion token corpus的统计

■ 基于 $p(k|.)$ 本身：因为water\fashion等词的干扰，不能很有效区分两词

■ 基于比值 $p(k|\text{ice})/p(k|\text{steam})$ ：可以更明显区分两词，因为干扰词的比例都为1

■ 因此，为学到更有区分性的向量，从同现概率的比例出发构建模型 \rightarrow

■ 后面要用到的一些记号

■ $p_{ij} = p(j|i) = X_{ij}/X_i$: 词 w_j 出现在词 w_i 的上下文中的概率

■ 其中

■ X_{ij} : 词 w_i 的上下文中出现词 w_j 的次数

■ X_i : 词 w_i 的上下文中所有词的次数

■如下演化过程(方式类似logistic回归的一种演化推导):

■ $F(w_i, w_j, \widetilde{w_k}) = \frac{p_{ik}}{p_{jk}}$ (直接用 $w_i, w_j, \widetilde{w_k}$ 表示对应词的词向量) \rightarrow

■ $F(w_i - w_j, \widetilde{w_k}) = \frac{p_{ik}}{p_{jk}}$ \rightarrow

■ $F((w_i - w_j)^T \widetilde{w_k}) = F(w_i^T \widetilde{w_k} - w_j^T \widetilde{w_k}) = \frac{p_{ik}}{p_{jk}}$ \rightarrow

■ $F(w_i^T \widetilde{w_k}) = \exp(w_i^T \widetilde{w_k}) = p_{ik} = \frac{X_{ik}}{X_i}$ \rightarrow

■ $w_i^T \widetilde{w_k} = \log(p_{ik}) = \log(X_{ik}) - \log(X_i)$ \rightarrow

■ $w_i^T \widetilde{w_k} + \log(X_i) = \log(X_{ik})$ \rightarrow

■ $w_i^T \widetilde{w_k} + b_i + \widetilde{b_k} = \log(X_{ik})$ \rightarrow

■ $J = \sum_{i,k=1}^V (w_i^T \widetilde{w_k} + b_i + \widetilde{b_k} - \log(X_{ik}))^2$

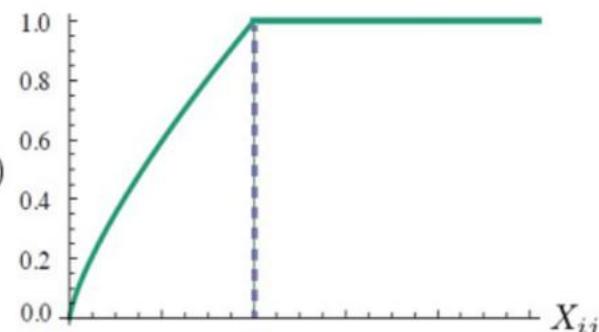
$$J = \sum_{i,k=1}^V (w_i^T \tilde{w_k} + b_i + \tilde{b_k} - \log(X_{ik}))^2$$

- 但是这个损失函数对每个词同等要求，各个词的词不同，作用也不同，为此设计加权重 $f(X_{ik})$ ：
- 设计权重函数满足：

- 1. $f(0) = 0$
- 2.频次高的权重大，因此要非递减的
- 3.频次太高的权重要控制，否者降低了其他词的作用

- 有很多，一个选择：

$$\blacksquare f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$



$$J = \sum_{i,k=1}^V (w_i^T \tilde{w_k} + b_i + \tilde{b_k} - \log(X_{ik}))^2$$

■ 损失函数最终为：

$$\blacksquare J = \sum_{i,k=1}^V f(X_{ik})(w_i^T \tilde{w_k} + b_i + \tilde{b_k} - \log(X_{ik}))^2 \quad (8)$$

■ 算法的计算复杂性：

■ 最差情况： $\mathcal{O}(|V|^2)$ 。

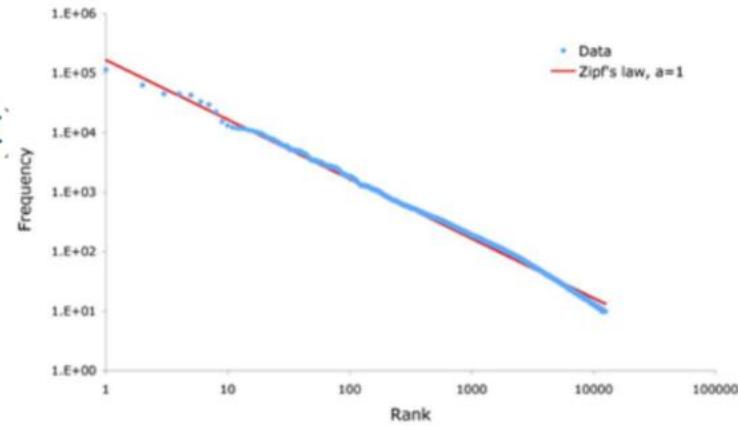
■ 基于：词出现的幂律分布 (Zipf律)

■ 获得复杂性：

$$\blacksquare |X| = \begin{cases} \mathcal{O}(|C|) & \text{if } \alpha < 1 \\ \mathcal{O}(|C|^{1/\alpha}) & \text{if } \alpha > 1 \end{cases}$$

■ 和 Skip-gram 等模型的关系

■ 参见 [Pennington 2014 EMNLP]



Zipf's law: 给定大规模语料, 词的出现频次和它在频次表中的排位成反比: $f = \frac{1}{r^\alpha}$

■实验结果比较 (表均来[Pennington2014EMNLP])

Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	<u>67.5</u>	<u>54.3</u>	<u>60.3</u>
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	<u>64.8</u>	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	<u>80.8</u>	61.5	<u>70.3</u>
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW [†]	300	6B	63.6	<u>67.4</u>	65.7
SG [†]	300	6B	73.0	66.0	69.1
GloVe	300	6B	<u>77.4</u>	67.0	<u>71.7</u>
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	<u>81.9</u>	<u>69.3</u>	<u>75.0</u>

词类比任务(accuracy)

Model	Size	WS353	MC	RG	SCWS	RW
SVD	6B	35.3	35.1	42.5	38.3	25.6
SVD-S	6B	56.5	71.5	71.0	53.6	34.7
SVD-L	6B	65.7	<u>72.7</u>	75.1	56.5	37.0
CBOW [†]	6B	57.2	65.6	68.2	57.0	32.5
SG [†]	6B	62.8	65.2	69.7	<u>58.1</u>	37.2
GloVe	6B	<u>65.8</u>	<u>72.7</u>	<u>77.8</u>	53.9	<u>38.1</u>
SVD-L	42B	74.0	76.4	74.1	58.3	39.9
GloVe	42B	<u>75.9</u>	<u>83.6</u>	<u>82.9</u>	<u>59.6</u>	<u>47.8</u>
CBOW*	100B	68.4	79.6	75.4	59.4	45.5

词相似任务(Spearman rank correlation)
词向量均为300维

■词向量的维数

■尝试、经验确定

■On the Dimensionality of Word Embedding, NIPS2018

用SG 任务	WordSim353	MTurk771	Google analogy
实验选择	56	102	220
方法的理论选择	+10% interval [48, 269]	+5% interval [67, 218]	+10% interval [48, 269]

用GloVe 任务	WordSim353	MTurk771	Google analogy
实验选择	220	860	560
方法的理论选择	+10% interval [160,1663]	+5% interval [290,1286]	+5% interval [290,1286]

■问题与发展

■每次用局部上下文进行预测，全局统计信息？

■suffer from the disadvantage that they do not operate directly on the co-occurrence statistics of the corpus. Instead, these models scan context windows across the entire corpus, which fails to take advantage of the vast amount of repetition in the data. From [Pennington2014EMNLP]

■低频词的词向量学习？

■关注低频词的词表示

- Turian等人发现：基于Collobert的分布式词表示与基于布朗聚类的词表示在高频词上表现类似(命名实体识别任务)，但是在低频词上Collobert 的词表示表现更差。
- Chen等人在2015年提出在中文上利用汉字作为特征来提升低频词的词表示，并得到一个字向量和词向量联合训练的词表示模型。
- Bojanowski等人在2016年利用英文这类具有丰富词形态学变化的语言的特点来提升低频词的词表示
-
- 超大规模模型及数据：预训练模型 (Transformer之后讲)

■问题与发展

■每次用局部上下文进行预测，全局统计信息？

■suffer from the disadvantage that they do not operate directly on the co-occurrence statistics of the corpus. Instead, these models scan context windows across the entire corpus, which fails to take advantage of the vast amount of repetition in the data. From [Pennington2014EMNLP]

■低频词的词向量学习？

■多义词的词向量学习？

■关注多义词的词表示

■Huang利用全局文本信息和局部信息来得到多义词的多个词向量。

■Eric H Huang, Richard Socher, Christopher D Manning, et al.

Improving word representations via global context and multiple word prototypes. ACL2012.

■Liu等人利用主题模型来给语料中的每个词标记上一个主题。然后模型根据词和主题来训练带主题的词表示。这样上下文词表示可以更灵活的获取，同时模型可以获得更好的文档表示。

■Yang Liu, Zhiyuan Liu, Tat-Seng Chua, et al. Topical word embeddings[A]. AAAI2015.

■ Matthew等人提出ELMo

- 基于语言模型训练获得词表示
 - 上下文化的词向量：某词的词向量是该词所有句子的函数，因此同一个词在不同句子(上下文)中获得的词向量也是不同的。也称为动态词向量，之前的词向量称为静态词向量(static)
 - 缓解多义问题，帮助多个NLP任务的性能提升
-
- ELMo也标志着不仅仅是预训练词向量，还有预训练语言模型的出现，即不仅仅是用大规模数据获得可用于后续任务的词表示，而是可以基于大规模数据获得多层次的语言信息表示，用于后续任务。
 - 后续Attention机制引入进一步产生了基于Attention的Transformer网络以及BERT等更强大的预训练语言模型，其中也包含对词的动态向量表示，在之后相应地方再进行介绍。

[Matthew2018 NAACL-HLT]Matthew E. Peters, et al. Deep Contextualized Word Representations. NAACL-HLT 2018.