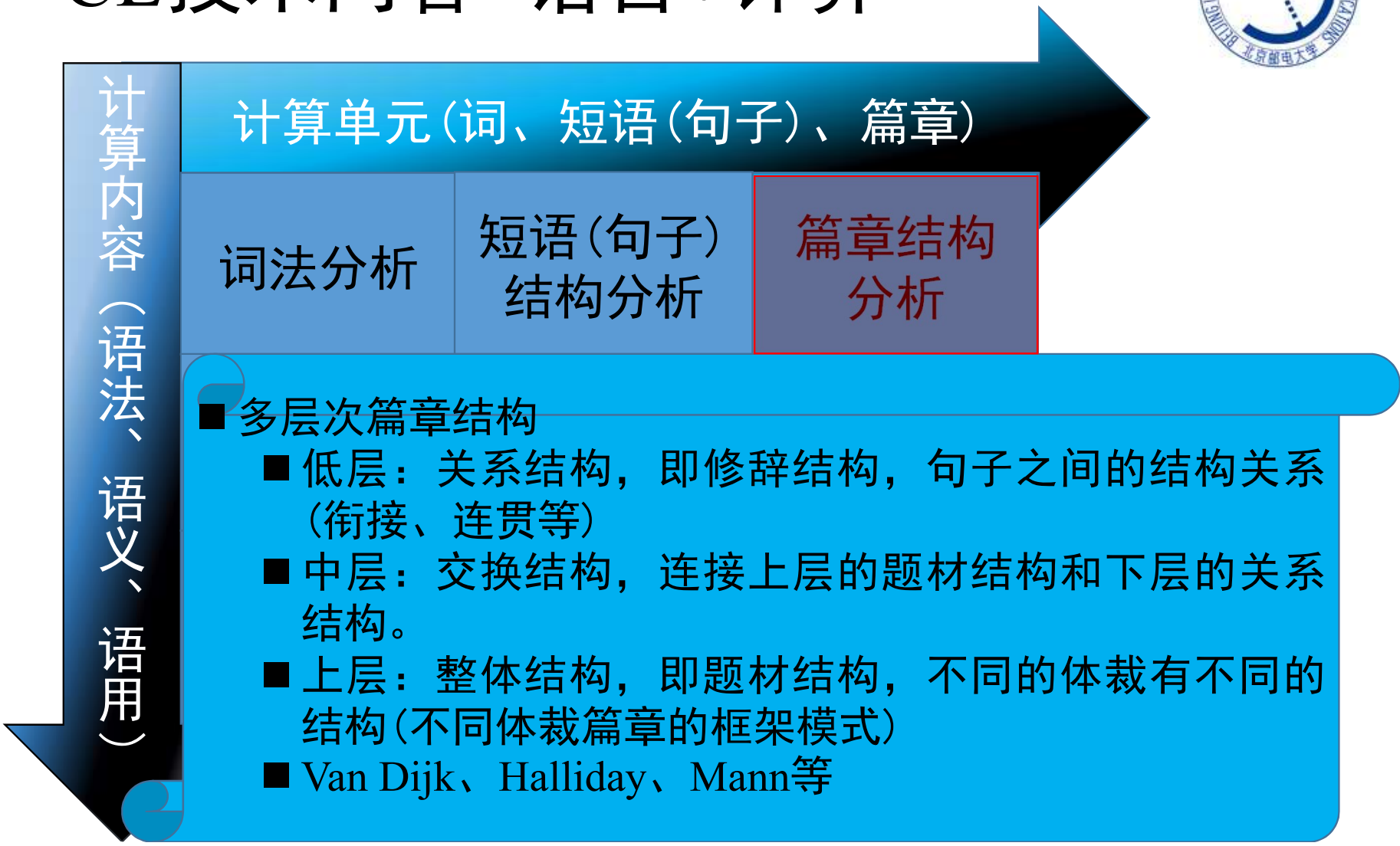
CL技术内容=语言+计算

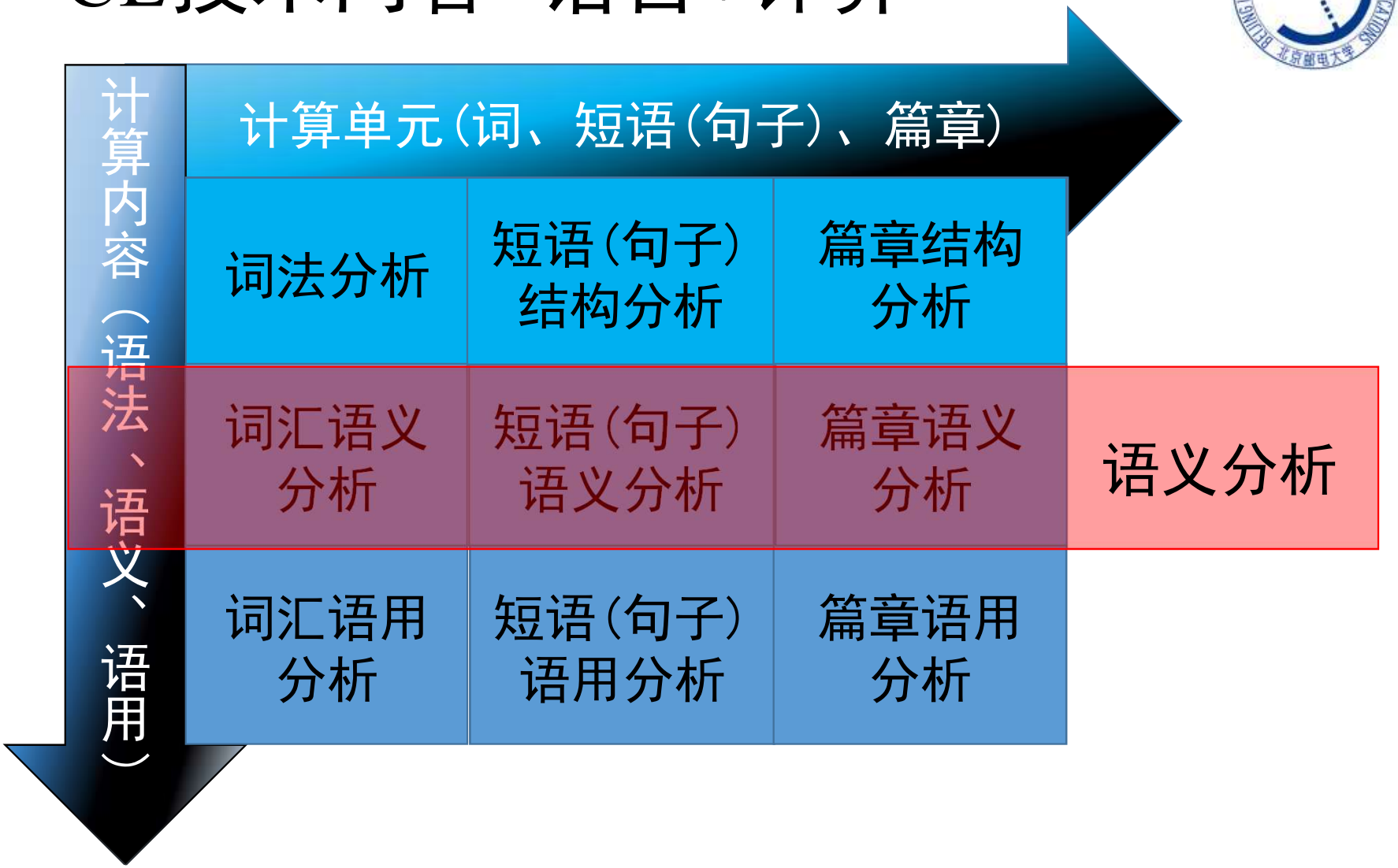
- 词汇的内部结构描述与分析
  - RG/FSA/FST
- 句子的内部结构描述与分析
  - 基于词汇的直接相依
    - N-gram：邻接词之间
    - Dependency parsing：隐结构依赖
  - 基于隐变量(POS\Chunk…)的相依：HMM…
  - 基于隐结构(NP\VP…)的相依：CFG…
  - 各有所长，各有所短

CL技术内容=语言+计算



CL技术内容=语言+计算

CL技术内容=语言+计算



CL技术内容=语言+计算

几个词

- 词典中：
- 空气： (现代汉语词典)
  - [1]构成地球周围大气的气体。
  - [2]气氛：学习空气浓厚

- Sense：义项
- Meaning：意义
- Semantics：语义

词义定义

- 基于符号的方法：
  - 用基本词定义：语义结构　（词的内部）
  - 用其他词定义：词间关系　（词词之间）
  - 结合的方法
- 基于向量的方法：
  - 同现词的高维向量
  - 低维实质向量

词义定义

- 基于符号的方法：
  - 用基本词定义：语义结构
  - 用其他词定义：词间关系
  - 结合的方法
- 基于向量的方法：
  - 同现词的高维向量
  - 低维实质向量

■基于语义结构的词义：一个词的词义是由若干个不可进一步进行意义分解的语义特征决定的。

词	语义特征		
父亲	+长辈	+男性	+可数
母亲	+长辈	+女性	+可数
水	+液体	+透明	-可数
笔	+固体	-透明	+可数

- 语义特征： 词义的基本单元，不可再分
  - 语义属性(Semantic properties)
  - 语义原语(Primitive)
  - 义原(Primitive)

- 问题：如何定义基本语义单元

- (语言学)专家分析决定
  - 语义单元可以是一些词，也可以是一些基础概念(义原、原语、…)
- 例子1
  - HowNet：基于义原进行词汇定义
    - 批准：DEF={ExpressAgreement|示同意}
    - 打1：DEF=buy|买
    - 打2：DEF=weave|辫编
    - 救灾DEF=rescue|救助,StateIni=unfortunate|不幸
    - HowNet(http://www.keenage.com/)

- 打
  - 打1：DEF=buy|买
  - E_C=~酱油，~张票，~饭，去~瓶酒，醋~来了
  - 打2：DEF=weave|辫编
  - 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) ？
- 进一步归结到词汇(或义原)间的关系

- 例子2
- Oxford学生词典
  - 定义2000基础词，其他词均由基础词定义

- 例3：一个方法
  - 1基于一个原则：原语不能由其他词所定义
  - 2利用词典来找到基本单元
  - 例子
    - 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 为一个基本单元

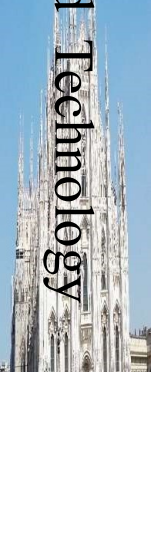
- 语义结构的心理学支持: 概念的特征表说
  - 概念由两部分组成：
    - 一些特征: 一个特征就是一个语义原语
    - 一些规则：如何结合不同特征的规则
- 例子:
  - 概念 “红圆”
    - 特征集 = {“红”, “圆”}
    - 规则： 特征”and” 操作

- 语义特征与语言使用有关联
- 语义特征可以更细致地约束词的使用
  - 如果某个词有一个语义特征是“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，也可作为语义特征抽取的一种途径
  - 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

- 语义结构分析的问题
  - 总共有多少语义特征：足够用 且 最少数量
  - 如何获得语义特征
    - 专家、新词、词义发展
    - 人工概念相对容易、自然概念比较难

词义定义

- 基于符号的方法：
  - 用基本词定义：语义结构
  - 用其他词定义：词间关系
  - 结合的方法
- 基于向量的方法：
  - 同现词的高维向量
  - 低维实质向量



■词间关系，先看义项间(单义词)关系

两个义项间的关系

- 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

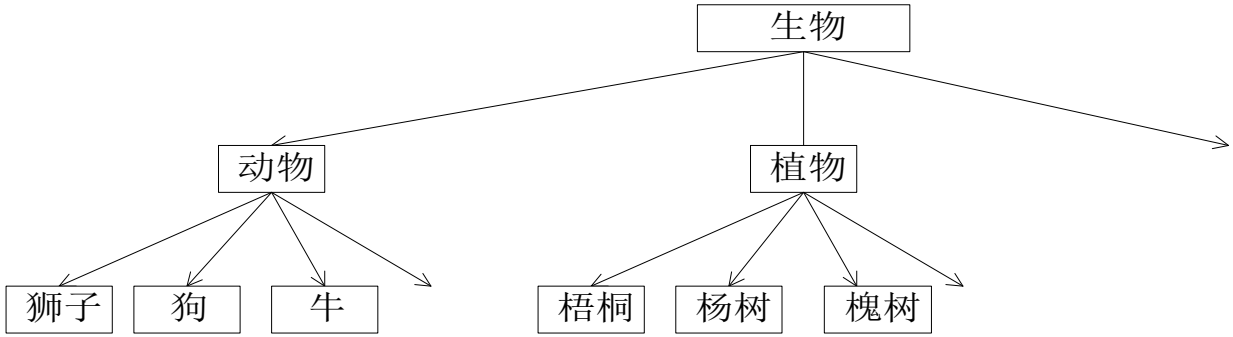


图 4-1 一个简单的分类体系

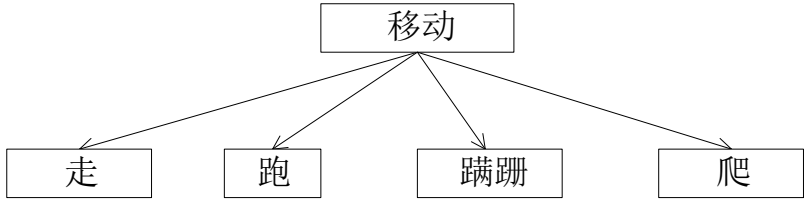


图 4-2 表示动作的义位构成上下义关系

- 全体-成员关系
  - 从全体到成员关系 (Has-Member)：：
    - 学院→系
  - 从成员到全体: ：
    - 系→学院

- 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): 具有相同或相近意义的不同词.
  - 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/reveive)，买/卖(buy/sell),…
  - 自反
    - Cleave(to split apart | to cling together)，无价?

- 换喻 (Metonyms)
  - 用对象的属性或某个侧面来指称对象
    - 中南海、克里姆林宫、白宫：用政府所在的建筑来指代政府

- 多义词(Polysemy)：一个词有多个义项
  - 意思：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)
  - 不同义项间有系统性关联 vs. 没有关联
- 省：节约;行政单位 二者无关
- 看：视、阅读、认为 有关，且：系统性相关➔
- 听、闻、触、嗅…

■虽然一个词可以有多个义项，但是在一个信息足够充分的特定上下文中，它只能取其中一个，让计算机自动完成词任务，就是词语义分析的一个基本任务：词义消歧(Word Sense Disambiguation:WSD)

词间相似性：比近义词更一般

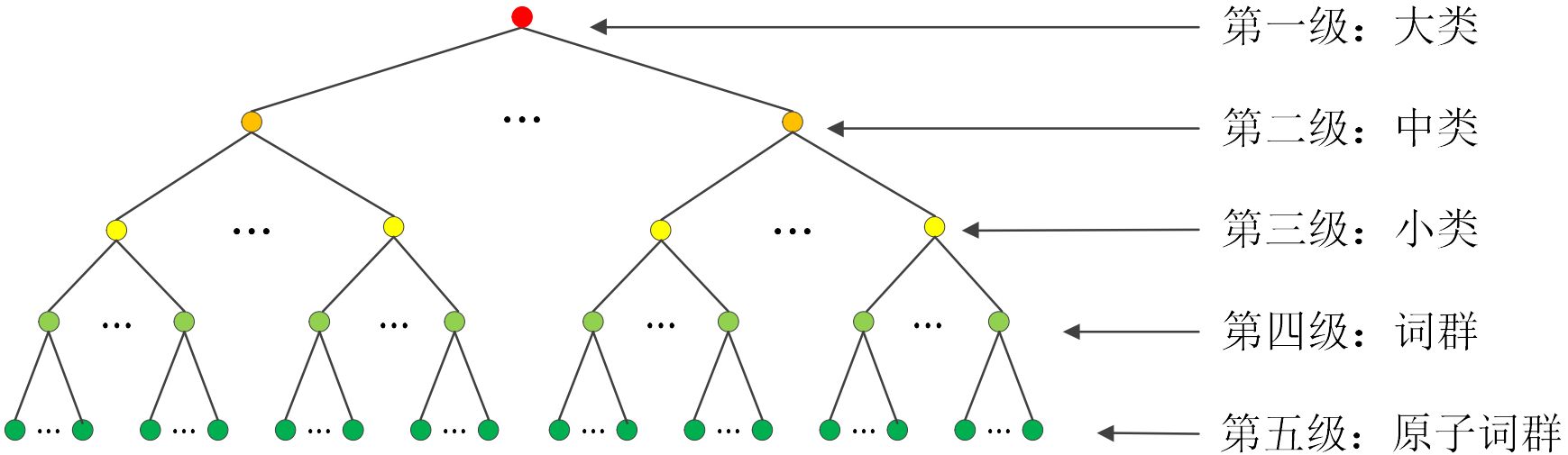
- sim(猫,狗)> sim(猫,桌子)
- sim(站,坐)> sim(站 vs 看待)
- 词间相似性具有很多重要语言使用和理解价值：
  - 例如：
  - 如果：sim(猫,狗)> sim(猫,桌子)
  - 那么：如果猫常和叫搭配使用，那么狗比桌子更可能和叫搭配使用。
- 如何度量相似性？是一个重要的语言处理问题

词间相关性

- 医生、病人、医院、急症
- 医生、毛巾、天空、经历
- 前者构成语义场：更可能共同出现在特定的场景中；
- 词间相关性对于消歧等任务都具有重要价值
- 主题模型有相通之处

■利用词间关系定义词义：

■同义词词林



■Aa01A01=人,士,人物,人士,人氏,人选

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	a	01	A	01

■Aa01A02=人类,生人,全人类

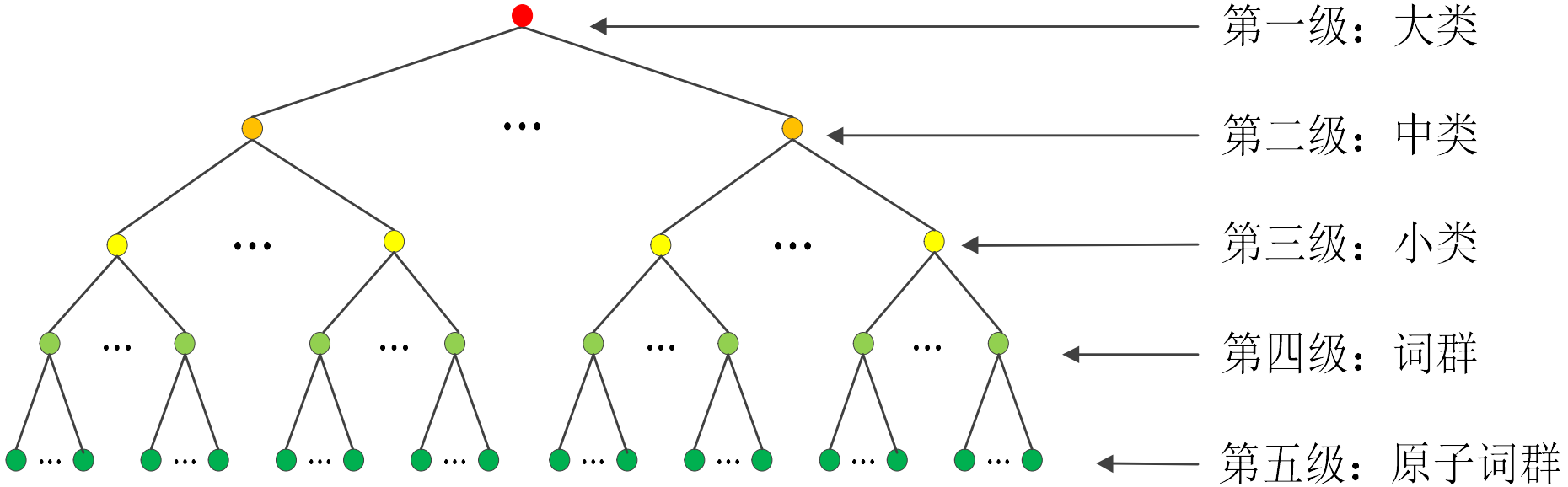
■Aa01A03=人手,人员,人口,人丁,口,食指

■Aa01A04=劳力,劳动力,工作者

■Aa01A05=匹夫,个人

■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：
  - 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.

词义定义

- 基于符号的方法：
  - 用基本词定义：语义结构
  - 用其他词定义：词间关系
  - 结合的方法
- 基于向量的方法：
  - 同现词的高维向量
  - 低维实质向量

- FrameNet：框架(frame)，例：
  - 框架：辨别 | Differentiation
    - 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

语义结构

相似、相关词

- 基于符号的词汇语义表达的问题
  - 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
- 直接进行数值化的表达?

词义定义

- 基于符号的方法：
  - 用基本词定义：语义结构
  - 用其他词定义：词间关系
  - 结合的方法
- 基于向量的方法：
  - 基于统计的高维向量及其降维
  - 基于预测的低维实质向量

- 基于数值向量的词表示
  - One-hot表示： |V|维向量，一词一维：(0,…,1,…,0)
  - 假设只有三个词：w1，w2，w3
  - w1=(1,0,0); w2=(0,1,0); w3=(0,0,1)
  - 无法表达词间关系的远近

- 基于上下文的表示(分布式distributional)：
  - 基本思想
    - 弗斯：观其伴，知其义
    - 基于目标词上下文的词来定义目标词
  - 方法：基于大规模语料的统计构建和词表维数一样的向量
    - 布尔型：|V|维向量，一词一维，该词是否出现：(1,0,1……)
    - 频次型： |V|维向量，一词一维，该词出现的次数：(10,0,7……)

- 例如,设有词表V={w1,..,wi,..wn}，对于词wi，设上下文窗口为K，则依据一个语料观测其中每个wi的上下文K个词，得到分布式表示：
  - 布尔型：n维向量 (0,…1,…1,…)：
    - 第j维为0：wj在上下文没有出现， 为1：wj在上下文没有出现。
  - 频次型： n维向量 (0,…10,…8,…)
    - 第j维为某个数值：对应wj在上下文窗口中出现过的次数

- 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

- 获得低维实值表示
  - 将高维向量降维
    - 选择某些维：tfidf、PMI…
    - 压缩到特定维：LSA、LDA…
  - 直接将词表示为低维实向量(Distributed)
    - 从概率神经语言模型开始的一系列工作…

词义定义

- 基于符号的方法：
  - 用基本词定义：语义结构
  - 用其他词定义：词间关系
  - 结合的方法
- 基于向量的方法：
  - 基于统计的高维向量及其降维
  - 基于预测的低维实质向量

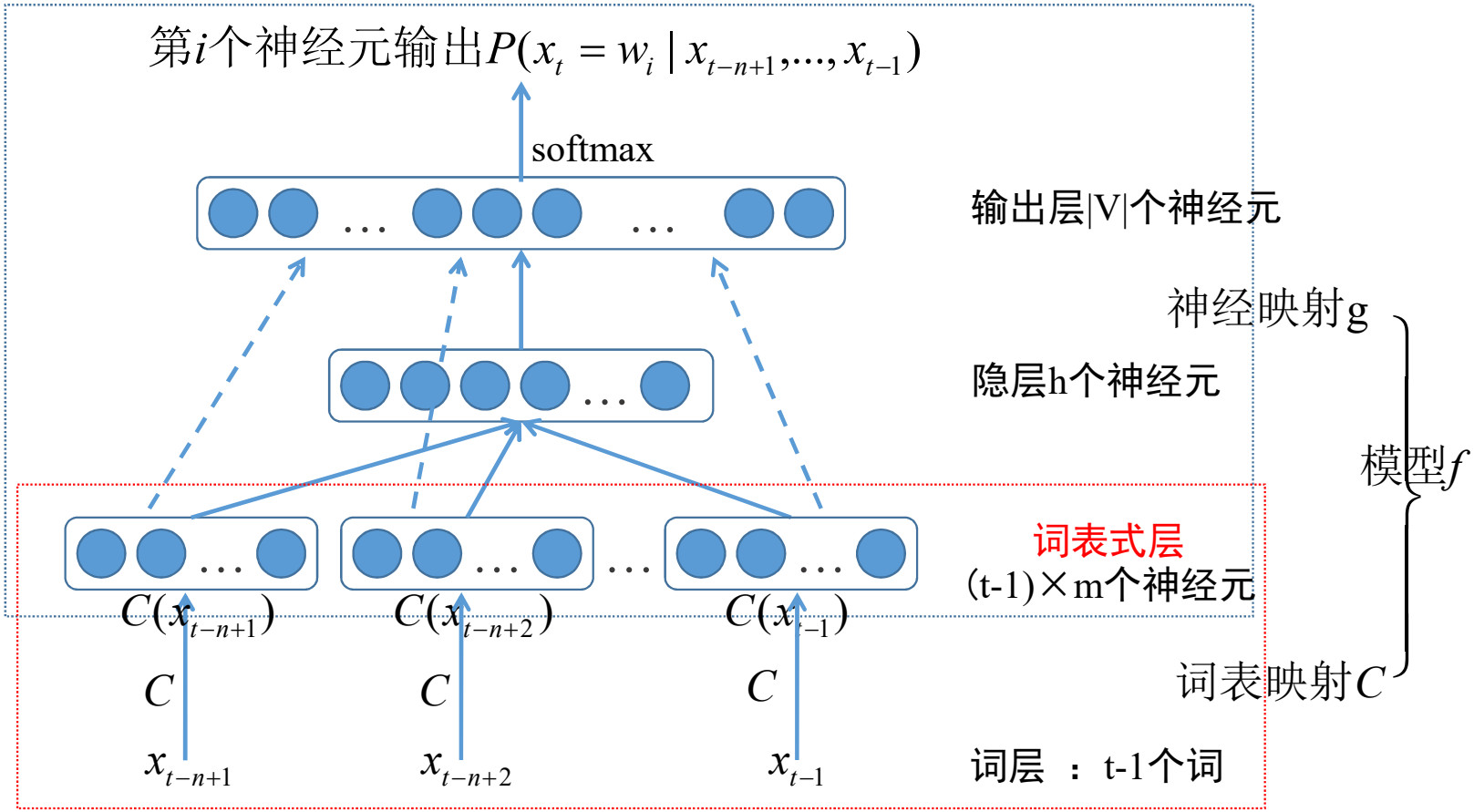
■术语

- ■Distributed Representation(分布式表示)
- ■Word Vector(Representation)(词向量(表示))
- ■Word Embedding (词嵌入表示)
- ■Continuous Space Representation (连续空间表示)

- 基于神经网络的词向量学习
  - 早期：思想、小规模尝试
  - 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.
  - D. E. Rumelhart, G. E. Hinton, R. J. Williams. Learning internal representations by back- propagating errors. Nature, 323:533.536, 1986.
  - 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
  - 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.
  - 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).
  - Turian, J., & Ratinov, L. (2010). Word representations : A simple and general method for semi-supervised learning. ACL2010 (pp. 384–394)
  - ……

■前述基于神经网络的语言语言模型中提到
■在获得语言模型的同时获得了词的分布式表示

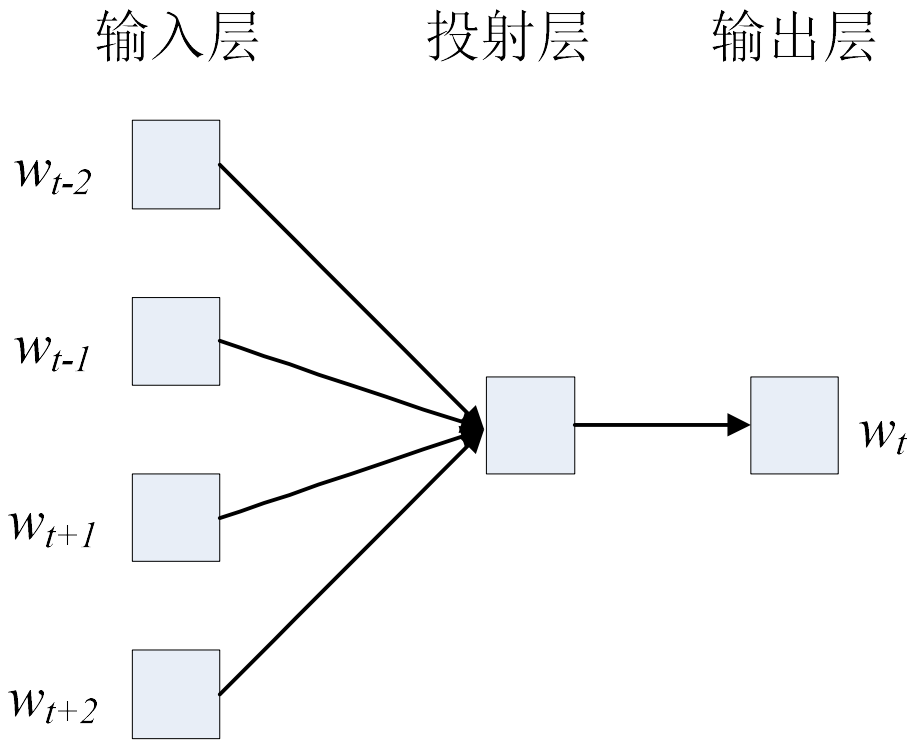


- 词向量作为语言模型训练的副产品
- 模型计算复杂性高
  - 隐层到输出层的联接
  - 输出层归一化求和
- 词表示获取也因此复杂
- 发展更为简洁、计算有效的方案

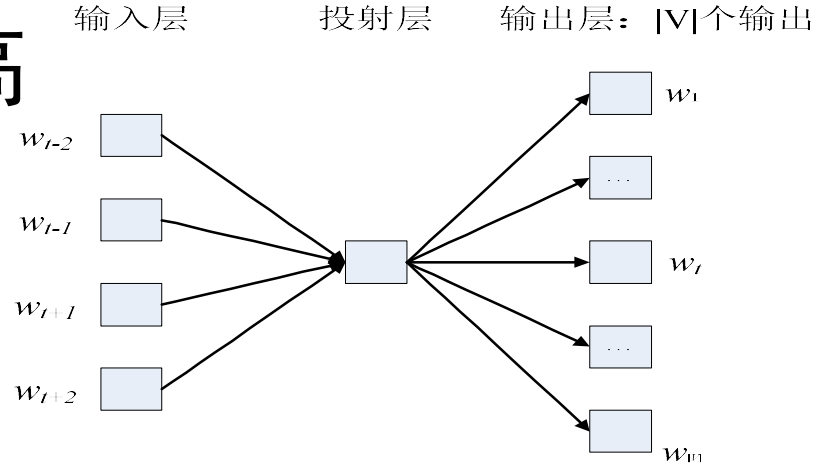
- 基于神经网络的词分布表示获取技术： 直接面向词向量学习
  - 训练复杂度更小、规模更大、更好的表示：推动了词表示的广泛应用
  - Tomas Mikolov, Greg Corrado, Kai Chen, Jeffrey Dean, Efficient Estimation of Word Representation in Vector Space, ICLR2013 workshop
  - Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems. 2013
  - 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

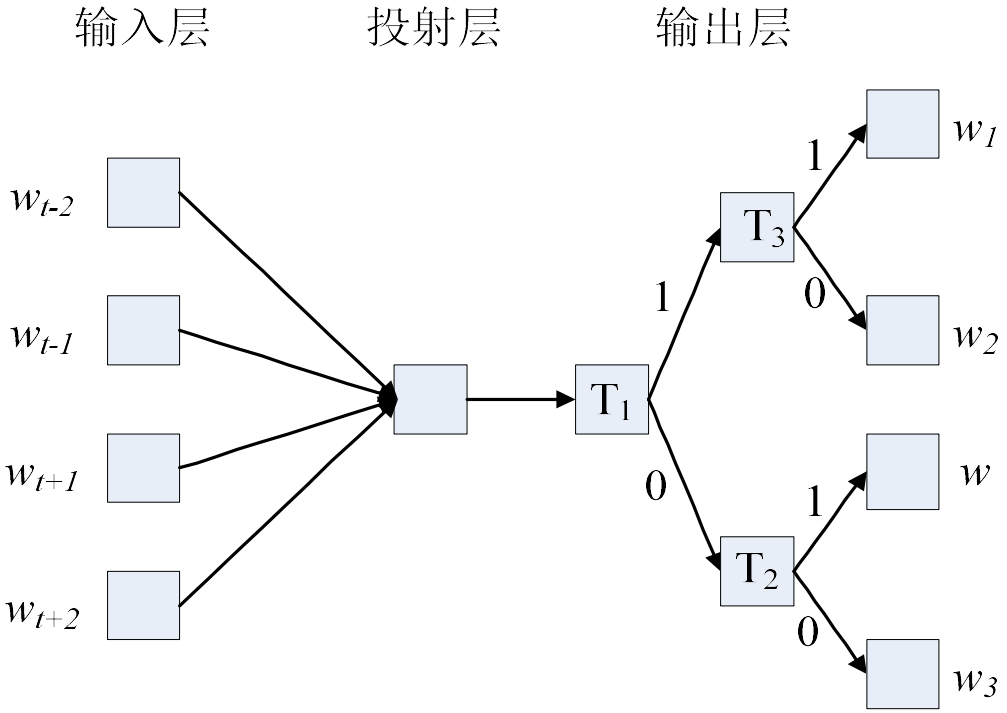
- CBOW(Continuous Bag-Of-Words)
  - 出发点：通过上下文的环境来预测当前词
  - 对于目标词w_t及其上下文$C(w_t)$：$P(w_t|C(w_t))$



- 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_t，C(w_t)：P(w_t|C (w_t))
- 理想目标极大P(w_t|C (w_t)),即P(w_t|C(w_t)>P(非w_t|C(w_t))
- 如何得到P(非w_t|C(w_t)?
- 一个直接的方案是构建|V|类的softmax，所有其他w_t均为非w_t，但是模型复杂度高
- 如何更高效？



■方案1：层次softmax



P(w1|context(w))=
p(T1=1/context(w))p(T2=1/context(w))

P(w2|context(w))=
p(T1=1/context(w))p(T2=0/context(w))

P(w|context(w))=
p(T1=0/context(w))p(T2=1/context(w))

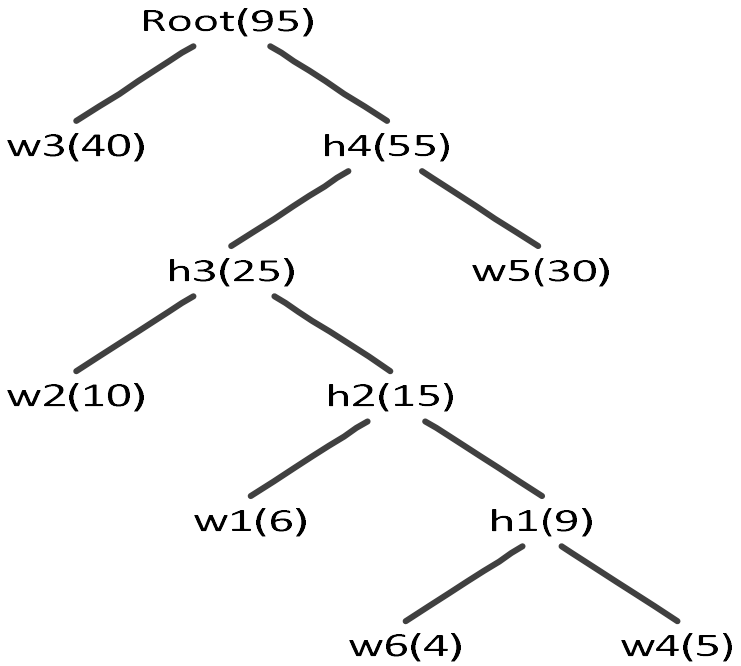
P(w3|context(w))=
p(T1=0/context(w))p(T2=0/context(w))

■将一个|V|分类分解为一系列的二分到达某个词

■问题：

- ■1）如何构造二分树
- ■2）多次二分：如何计算每条路径概率
- ■3）如何优化网络参数

- 词的哈夫曼树
  - 基于语料中的词频
  - w1: f(w1), w2: f(w2), ……, wn: f(wn)
  - 例 f(w1)=6,f(w2)=10,f(w3)=40,f(w4)=5,f(w5)=30,f(w6)=4

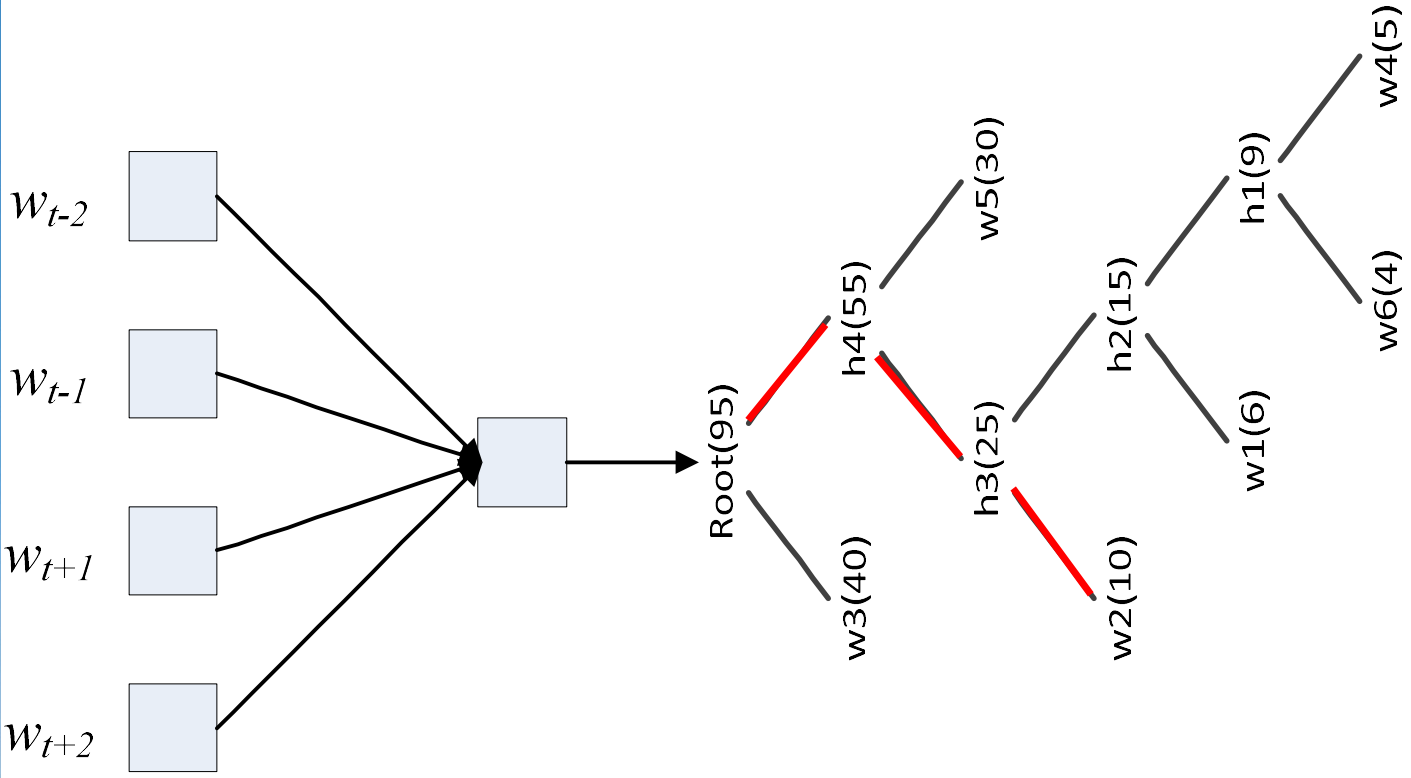


从根节点出发，频率高的词具有较短的路径

■层次softmax算法

■输出方为一个哈夫曼树，树的叶节点是词。

■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：要预测w2，那么，需要三次分类：

■1)先在root节点分到h4,

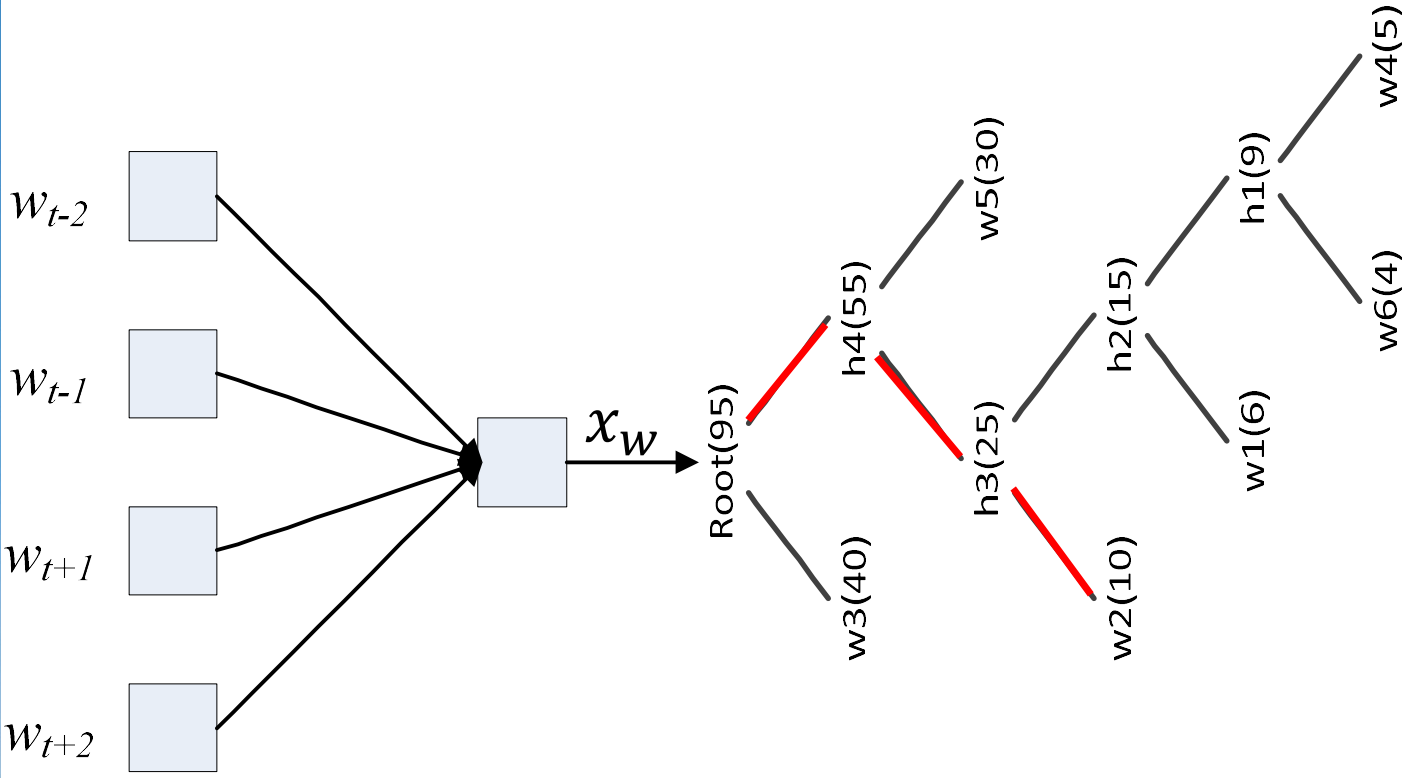
■2)再在h4节点分到h3

■3)最后在h3节点分到w2

■层次softmax算法

■输出方为一个哈夫曼树，树的叶节点是词。

■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：要预测w2，那么，需要三次分类：

■1)先在root节点分到h4,

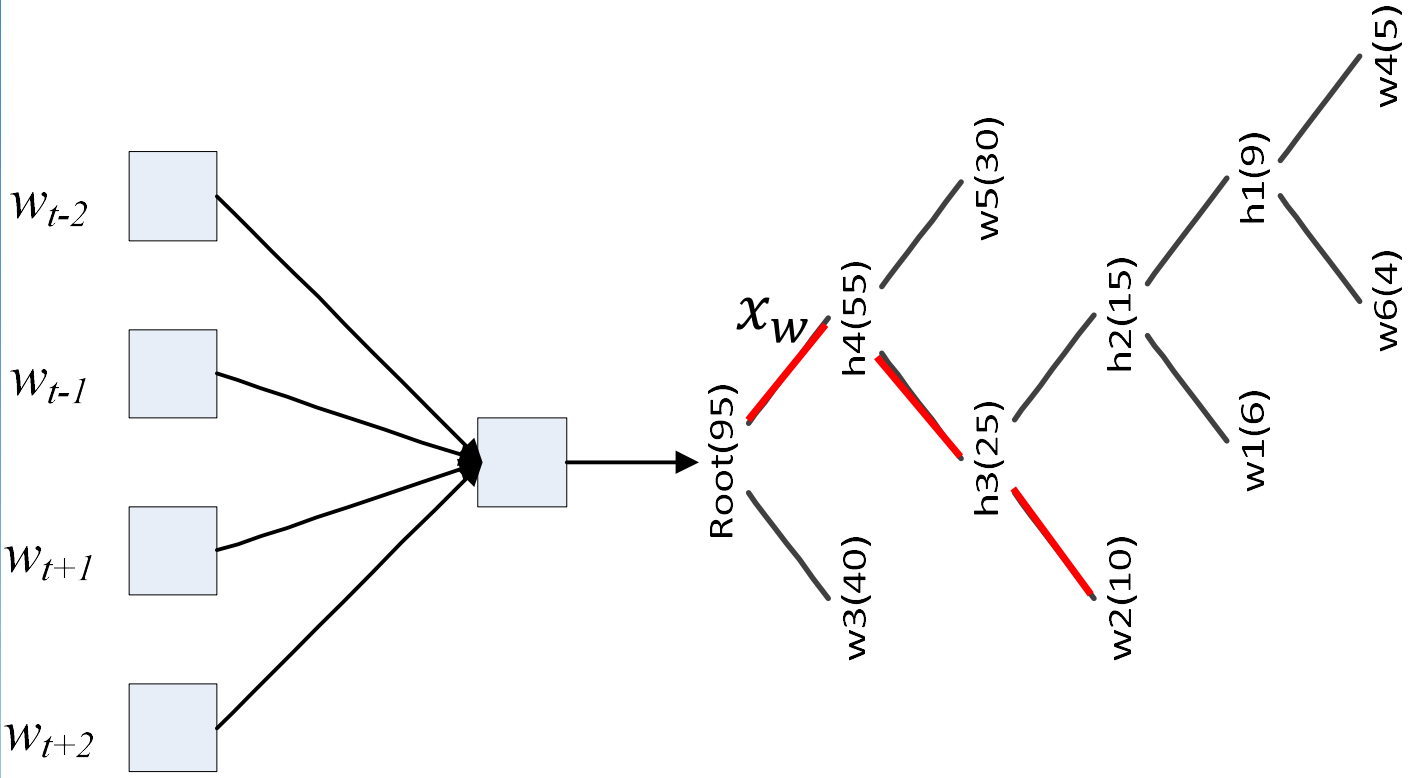
■2)再在h4节点分到h3

■3)最后在h3节点分到w2

■层次softmax算法

■输出方为一个哈夫曼树，树的叶节点是词。

■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：要预测w2，那么，需要三次分类：

■1)先在root节点分到h4,

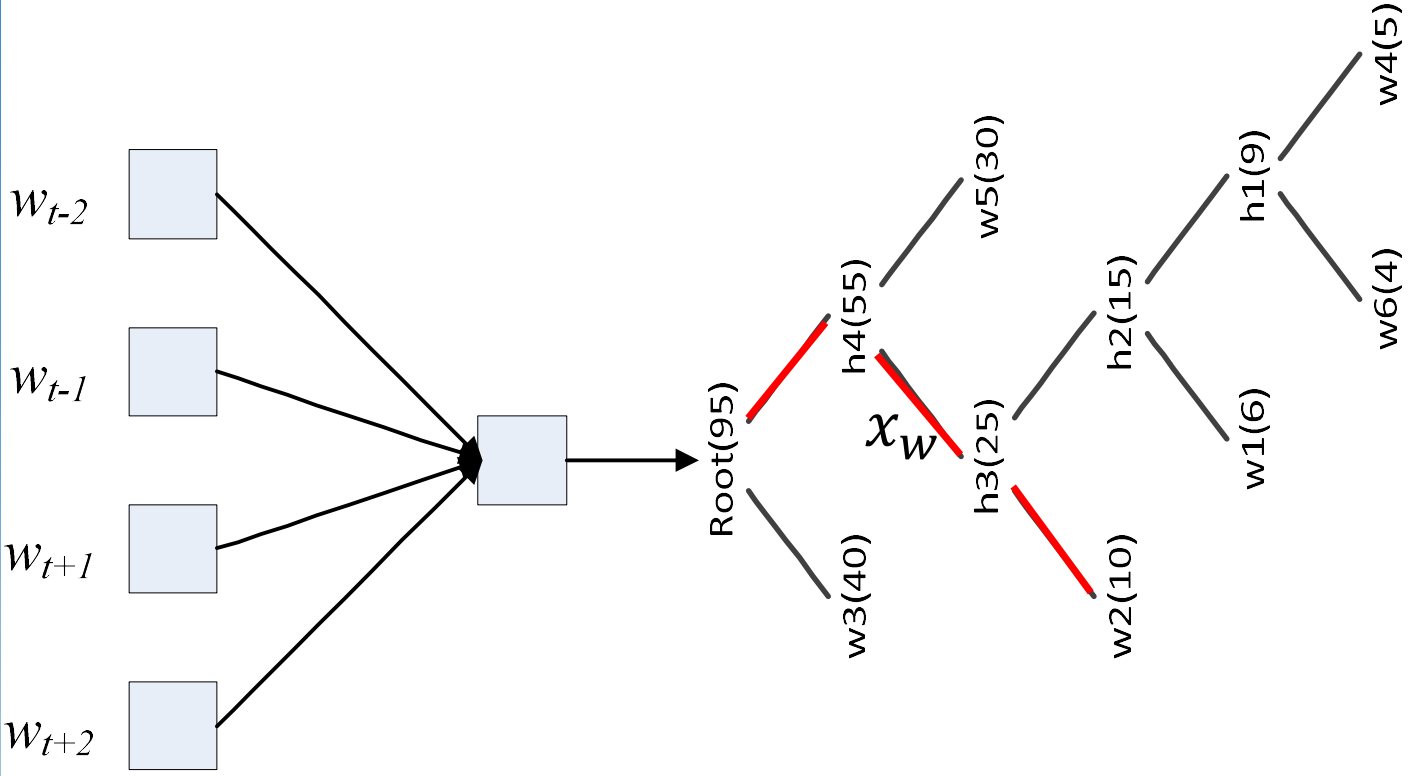
■2)再在h4节点分到h3

■3)最后在h3节点分到w2

■层次softmax算法

■输出方为一个哈夫曼树，树的叶节点是词。

■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：要预测w2，那么，需要三次分类：

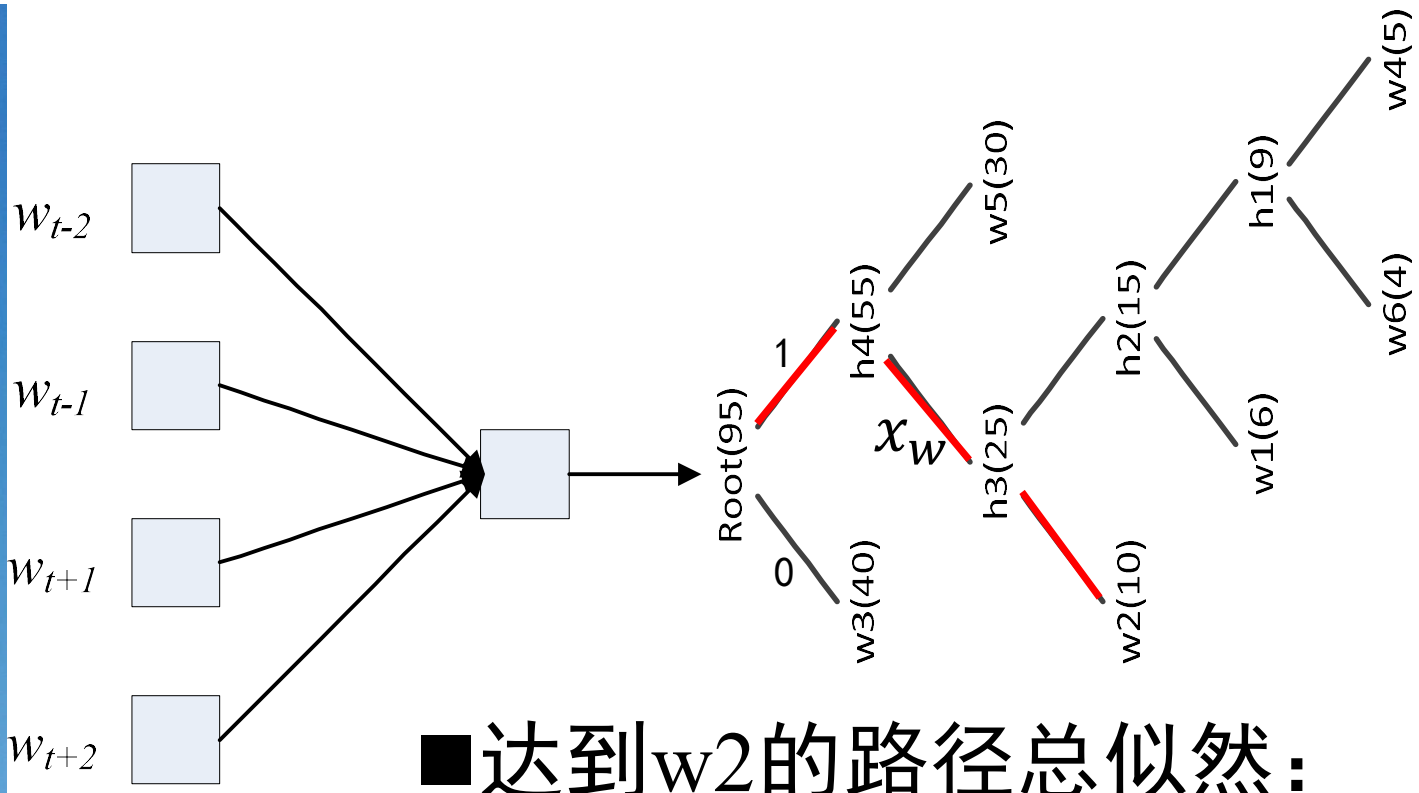
■1)先在root节点分到h4,

■2)再在h4节点分到h3

■3)最后在h3节点分到w2

■每个二分类(y={0,1})都采用逻辑斯特回归(左0右1)：

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



■达到w2的路径总似然：

■$[\sigma(x_w \cdot \theta_{root})]^{1-1}.\,[1-\sigma(x_w \cdot \theta_{root})]^{1}$
$*[\sigma(x_w \cdot \theta_{h4})]^{1-0}.\,[1-\sigma(x_w \cdot \theta_{h4})]^{0}$
$*[\sigma(x_w \cdot \theta_{h3})]^{1-0}.\,[1-\sigma(x_w \cdot \theta_{h3})]^{0}$

■目标：极大化上述似然

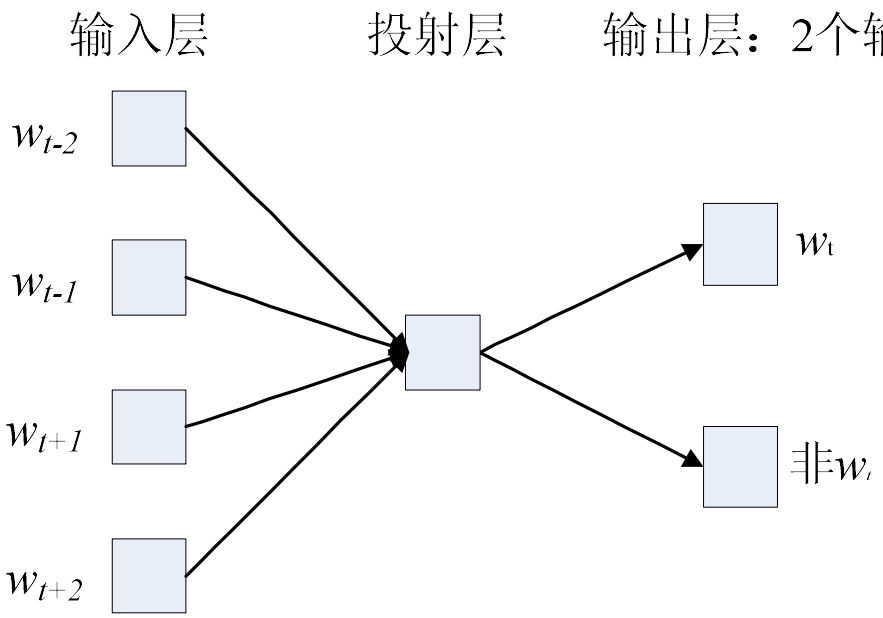
- 一般地，每个二分类的概率：
  - $P(d_j^w|x_w,\theta_{j-1}^w) =$

$$[\sigma\left(x_w^T\theta_{j-1}^w\right)]^{1-d_j^w} \cdot [1-\sigma\left(x_w^T\theta_{j-1}^w\right)]^{d_j^w}$$

  - 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。

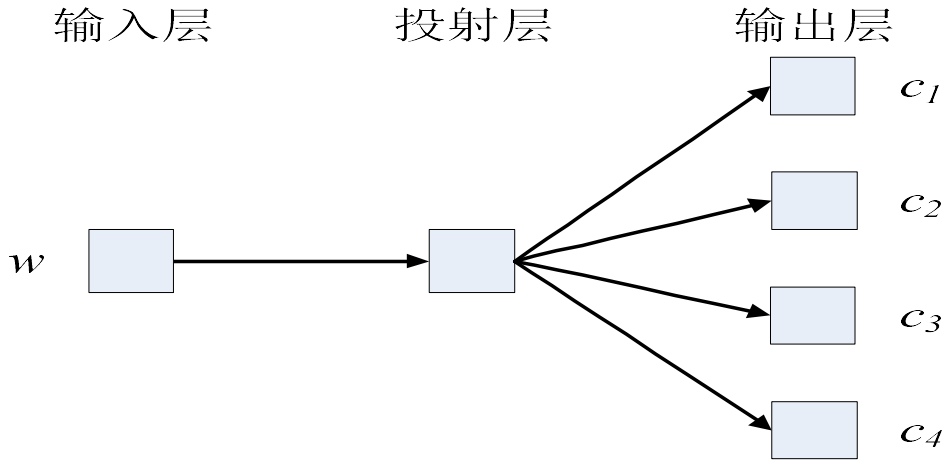
- 训练的目标就是极大化到达正确词的路径的概率
  - $P(\mathrm{w}|\mathrm{Context(w)}) = \prod_{j=2}^{l^w} P(d_j^w|x_w, \theta_{j-1}^w)$
    - $P(d_j^w|x_w, \theta_{j-1}^w) = [\sigma(\mathrm{x_w^T\theta_{j-1}^w})]^{1-d_j^w} \cdot [1-\sigma(\mathrm{x_w^T\theta_{j-1}^w})]^{d_j^w}$
  - 要使这个概率最大化可以采用随机梯度下降法，对参数x_w和θ_{j-1}^w进行更新
  - 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
- 能否直接解决用非w而不是用所有w？



- 负采样(Negative Sampling)方法
- 结合Skip-gram模型来介绍(SGNS)

- Skip-gram：
- 问题设定：预测当前词的上下文词



- 词w及其可能的上下文词$c \in C(w)$，构建参数化条件概率分布$\mathrm{p}(c|w;\theta)$,
- 最优参数是对于语料Text极大化似然：
  - $argmax_\theta \prod_{w \in Text} [\prod_{c \in C(w)} p(c|w;\theta)]$

- 进一步假设$p(c|w;\theta)$的基于词向量形式：
- 设$v_w, v_c \in R^d$分别为w, c的d维向量表示，定义softmax概率：
- $p(c|w;\theta) = \frac{e^{v_w * v_c}}{\sum_{c' \in C(w)} e^{v_w * v_{c'}}}$，（分母：配分函数）
- 参数θ就是v_w, v_c，总规模为$|C(w)| \times |V| \times d$
- 实际上，任何一个词都有可能是另一个词的上下文，因此，每个词都有作为上下文的词向量和作为目标词的词向量，前述目标函数变为：
- $argmax_\theta \prod_{w \in Text} [\prod_{c \in C(w)} \frac{e^{v_w * v_c}}{\sum_{c' \in C(w)} e^{v_w * v_{c'}}})]$

- 配分函数$\sum_{c\prime\in C(w)} e^{v_w * v_{c\prime}}$计算量大，转换为二分类问题，记：
  - $p(d=1|w,c)$为c是w的上下文的概率
  - $p(d=0|w,c)$为c不是w的上下文的概率，显然有
  - $p(d=0|w,c)$ =1- $p(d=1|w,c)$

- 同样用词向量以及softmax来参数化概率（逻辑斯特回归）为：

$$p(d=1|w,c;\theta)=\frac{e^{v_w * v_c}}{1+e^{v_w * v_c}}=\frac{1}{1+e^{-v_w * v_c}}=\sigma(v_w * v_c)$$

$$p(d=0|w,c;\theta)=1-\sigma(v_w * v_c)=\sigma(-v_w * v_c)$$

■则, 如果给定样本集$D=\{c,c$是w的上下文$\}$，$D'=\{c,c$不是w的上下文$\}$，参数优化目标是极大化似然：

■$argmax_\theta \prod_{c\in D} p(d=1|w,c;\theta) \prod_{c\in D'} p(d=0|w,c;\theta)$, 也即:

■$argmax_\theta (\sum_{c\in D} \log p(d=1|w,c;\theta) + \sum_{c\in D'} \log p(d=1|w,c;\theta))$

$=argmax_\theta \sum_{(w,c)\in D} \log\sigma(v_w * v_c) + \sum_{(w,c)\in D'} \log\sigma(-v_w * v_c)$

■极大化目标可以获得参数v_w、v_c，即词向量。

两个问题

1、样本构建

2、目标词还是上下文词的向量

■1、样本集构建

■设有语料片段：古 埃及 同时 步入 文明
c1 c2 w c3 c4

正例样本	负例
古	非“古”：文本，现在，未来，可以…
埃及	非“埃及”：学习…
步入	非“步入”：…
文明	非“文明”：…

■显然，负样例可以很多，因此：

■一个正例选几个负例?

■如何选择这些负样例？

- 一个正例选几个负例?
  - 数据小5-20，数据多2-5
  - 一个正例c对应多个负例c_1、…、 c_k构成一组样本，对应这组样本，极大化：

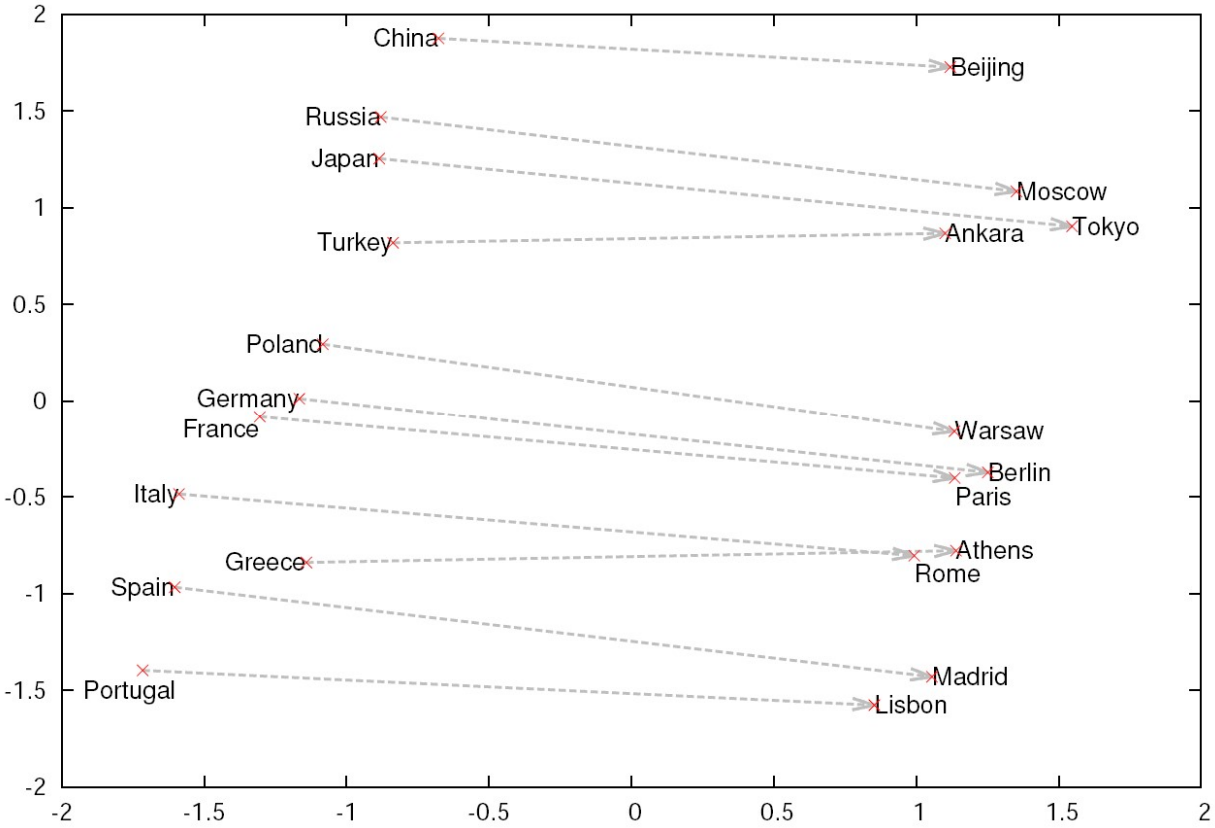
$$argmax_\theta[\log\sigma(v_w * v_c) + \sum_i \log\sigma(-v_w * v_{c_i})]$$

- 如何选择这些负样例？
  - $p_\alpha(\tilde{c}) = \frac{count(\tilde{c})^\alpha}{\sum_{\tilde{c}\prime} count(\tilde{c}\prime)^\alpha}$
  - 常用α=0.75：平缓不同词的采样概率
    - 提高低频词的采样率
    - 降低高频词的采样率

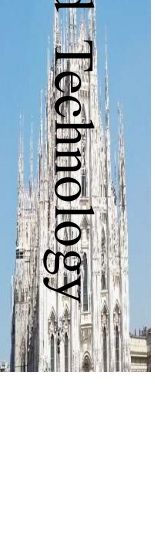
- 2、目标词还是上下文词和向量
- 一个词极可能是当前词w也可能是上下文词c，分别都可获得一个向量表示，哪个?
- 几种可能的选择：
  - 用作为当前词w的向量
  - 作为w的向量加上作为上下文词c的向量
  - 作为w的向量拼接作为上下文词c的向量，形成2d维的词向量
  - …

■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



- 词表示评估
- 外部任务：
- 内在任务：
  - 词相似度任务
    - 数据集：
    - 汉语：PKU500、CWE297,CWE240
    - 英语：WordSim353、RW、MEN、
    - 德语：ZG222、Gur250
  - 词类比任务
    - 北京:中国=巴黎:法国，CA8



■词向量的应用

- 词向量的应用
  - 简单地作为一种特征
  - 作为输入，但是固定
  - 作为输入，训练时也可调整(fine-tuning)

- 分布式词表示应用的研究：已用于几乎所有语言处理任务中
  - 用于切分、NER、POS
    - Xiaoqing Zheng et al. Deep Learning for Chinese Word Segmentation and POS Tagging, EMNLP2013
  - 用于Chunking、SRL
    - Collobert, R., Weston, J., & Karlen, M. (2011). Natural Language Processing (almost) from Scratch. JMLR
  - 用于Paraphrase Detection
    - Socher R., et al. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. ANIS 24
  - 用于Sentimental Analysis
    - Richard Socher et al. Semi-supervised recursive autoencoders for predicting sentiment distributions. EMNLP 2011
  - 用于Parsing
    - Socher, R., et al. Parsing natural scenes and natural language with recursive neural networks. ICML2011.
  - ……

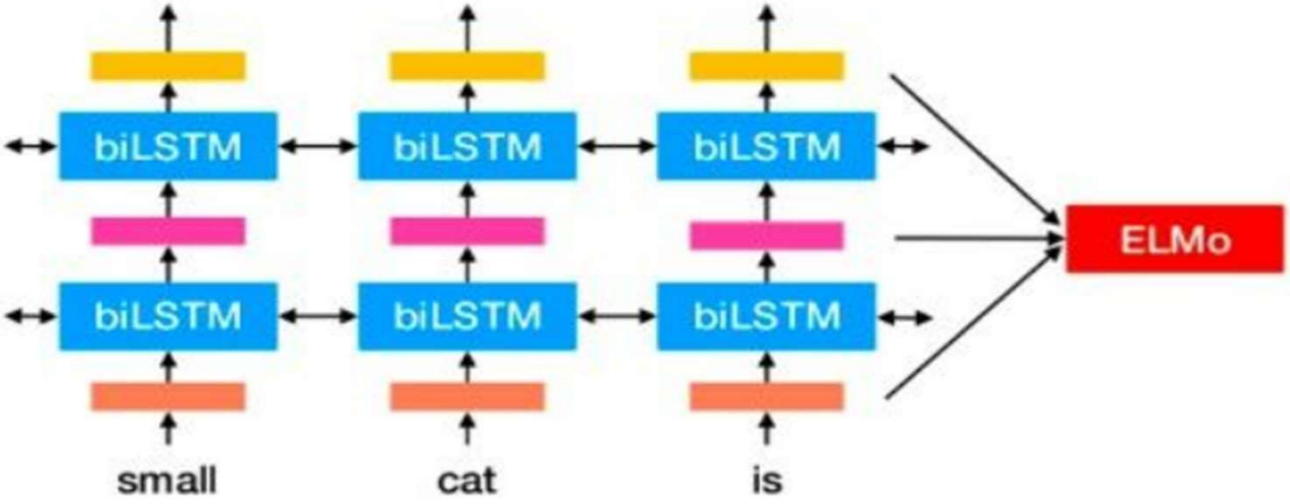
- 关于词向量的进一步研究
  - 词向量学到了什么？
  - 低频词的词向量学习？
  - 多义词的词向量学习？

- 基于矩阵分解的词表示构建
  - GLoVe：重点在构建词的共现矩阵，基于共现矩阵来进行矩阵分解

- 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FastText
  - ……

- 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.

- ELMo(Embedding from Language Models) 开始动态词向量
  - Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer: Deep Contextualized Word Representations. NAACL-HLT 2018.
  - 双向语言模型 + 上下文表示并入词



■缓解多义问题

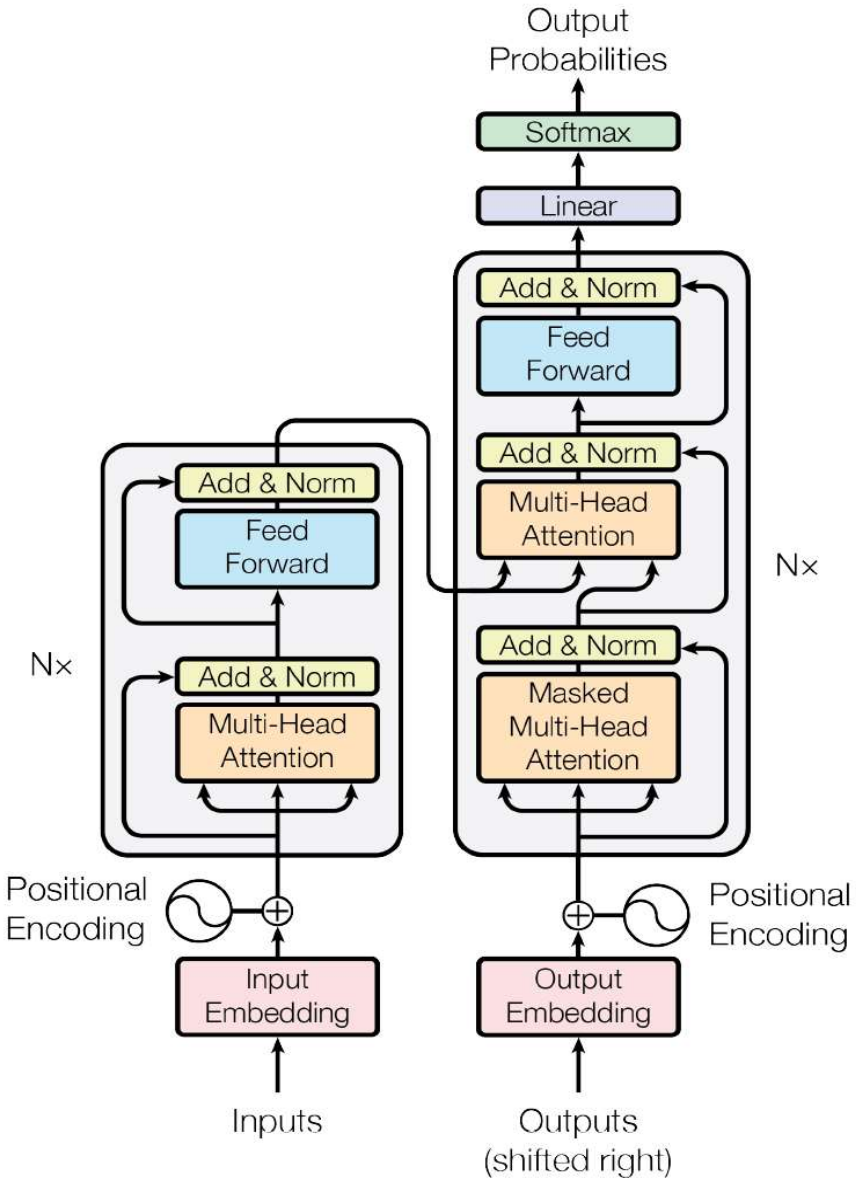
	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

■帮助多个NLP任务的性能提升

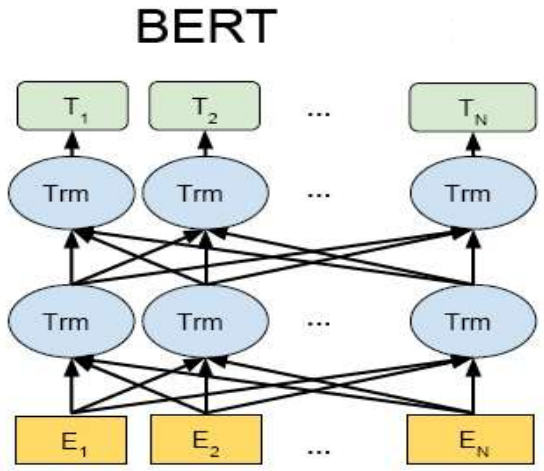
TASK	PREVIOUS SOTA		OUR BASELINE	ELMO + BASELINE
SQuAD	Liu et al. (2017)	84.4	81.1	85.8
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17
SRL	He et al. (2017)	81.7	81.4	84.6
Coref	Lee et al. (2017)	67.2	67.2	70.4
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5

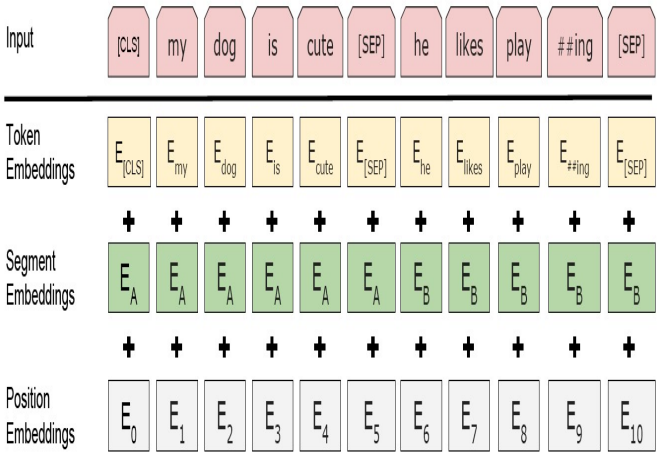
- Transformer:
  - Ashish Vaswani, et al. Attention is all you need. NIPS2017.

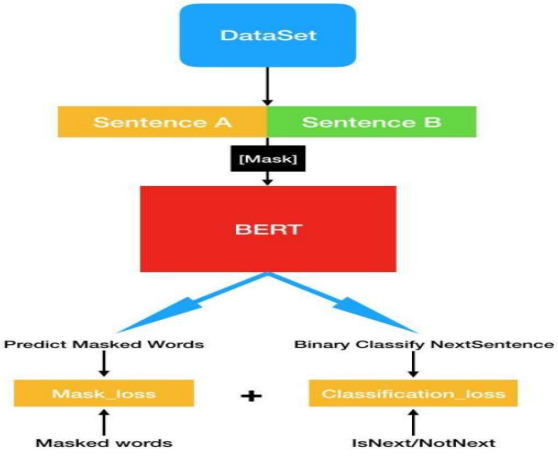
- 自注意
- 多头注意
- 更深的网络



- BERT(Bidirectional Encoder Representations from Transformers)
  - BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv 2018
  - 综合Transformer 和 ELMo
  - https://github.com/google-research/bert







■帮助多个NLP任务的性能提升

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$BERT_{BASE}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

■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]

- 基于SGNS的汉语词向量学习和评估作业
  - 详细说明见：SGNS作业说明.doc
  - 另外有数据文件： train.txt、 pku_sim_test.txt
- 提交时间：北京时间2020年12月20日24:00
- 提交内容：
  - 算法描述文档
  - python源码
  - 词相似度结果文本

谢谢！