大数据分析

课程总复习-Part2

程学旗

靳小龙

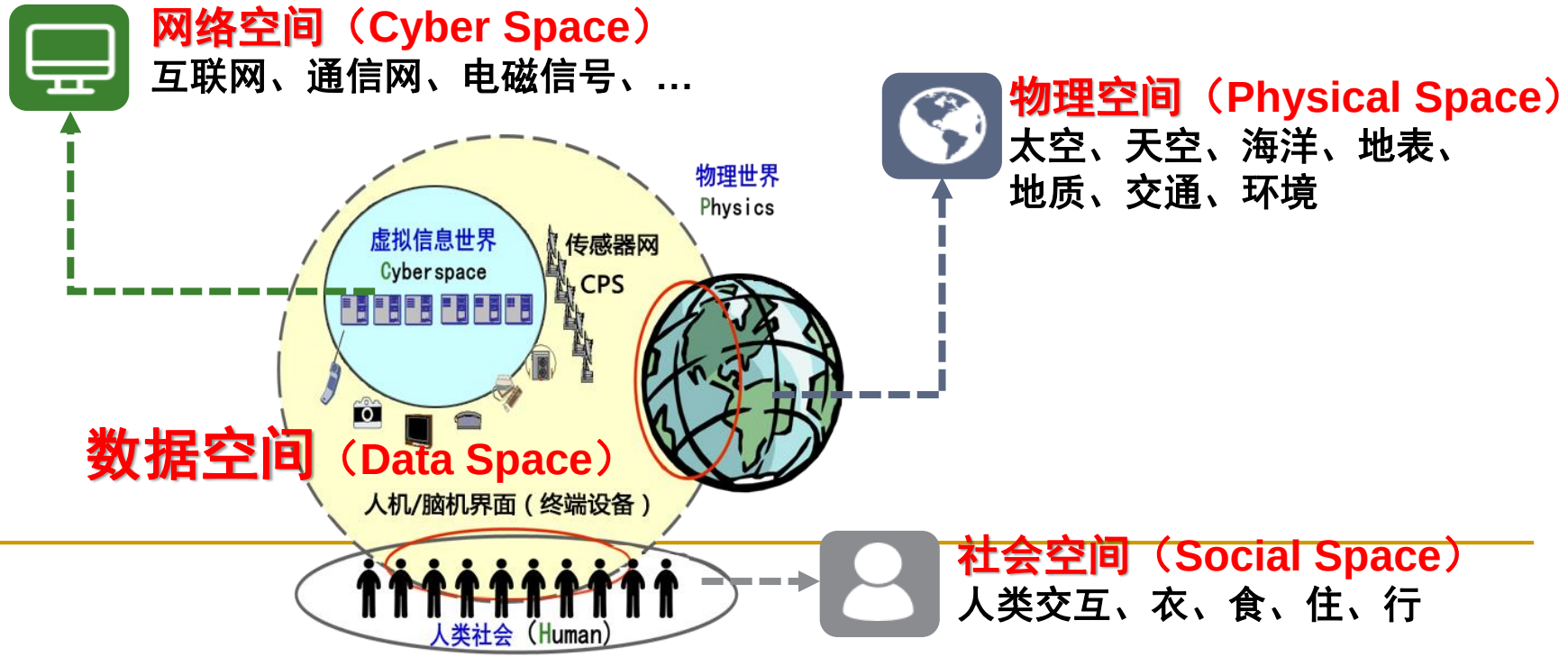
刘盛华

Outline

- **大数据与大数据分析简介**
- 大数据分析技术与系统
- 大数据统计分析
- 大数据机器学习
- 数据驱动的自然语言处理
- 文本大数据分析
- 知识图谱与知识计算
- 大图数据分析
- 社会媒体分析
- 数据与算法安全

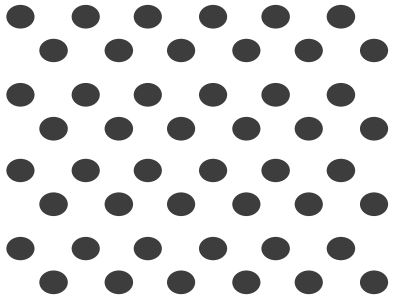
大数据的定义

- 在可容忍的时间内，无法用传统IT架构和软硬件工具对其进行全生命周期的感知、传输、存储、管理、计算和服务的数据集合；
- 大数据是信息世界(Cybernetics)、物理世界(Physical World)与人类社会(Human Society)三元世界彼此关联、动态交互的数字化、数据化呈现；



大数据的基本特征

Volume



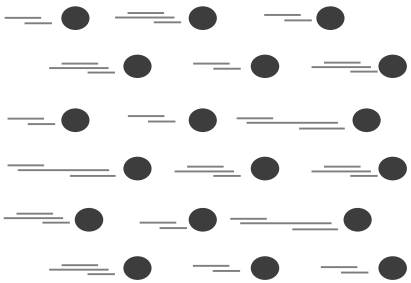
体量巨大

From terabytes to exabyte to zetabytes of existing data to process

IDC Analyze the Future

到2020年，数据总量达40ZB，人均5. 2TB

Velocity



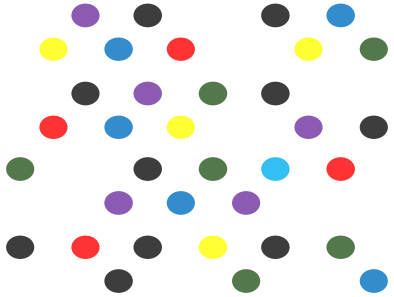
速度极快

Batch data, near-time data, real time data, streaming data, miliseconds to seconds to respond

facebook

分享的内容条目超过25亿个/天，增加数据超过500TB/天

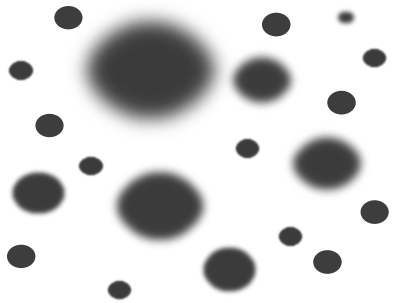
Variety



模态多样

Structured, semi-structured, unstructured, text, pictures, multimedia

Veracity



真伪难辨

Uncertainty due to data inconsistency & incompleteness, ambiguities, deception, model approximation

芦山地震十大不实谣言

2013年04月24日17:41　来源：人民网　手机看新闻

8. 地震局内部消息成都9. 2级地震

谣言：自称地震局内部人员的网民称发生9. 2级地震。”

大数据的大价值（Value）

大数据

数据采集/汇聚

- 面向人机物三元世界；
- 全面、准确、优质、高效；
- 多源异构、关联汇聚

数据存储/治理

- 将数据记录在计算机内部或外部存储介质，并进行安全与隐私保护等处理；

数据处理

- 检索、比对、排序、推荐、溯源、抽样、化简、统计等；

数据分析

- 分析数据所蕴含的模式、关系、结构、趋势、特征等；

数据应用

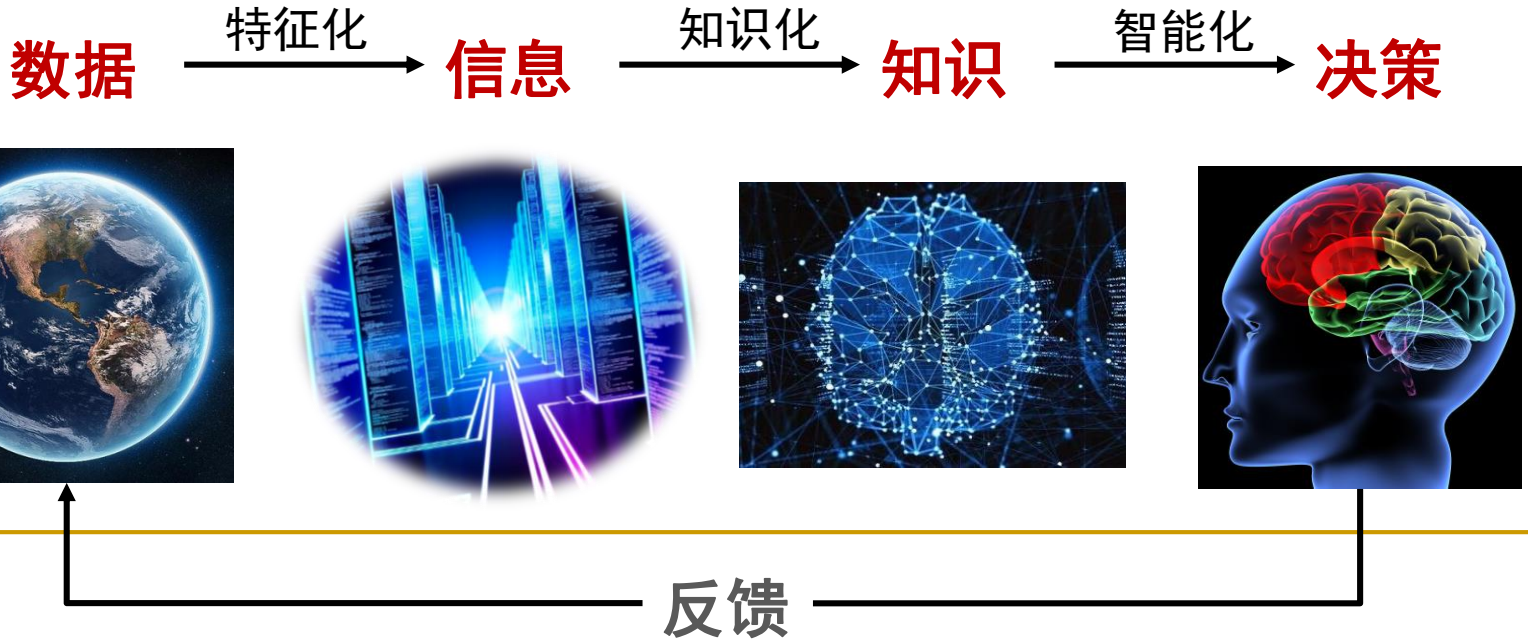
- 将数据分析结果与领域知识结合，形成决策并解决问题的过程；

大价值

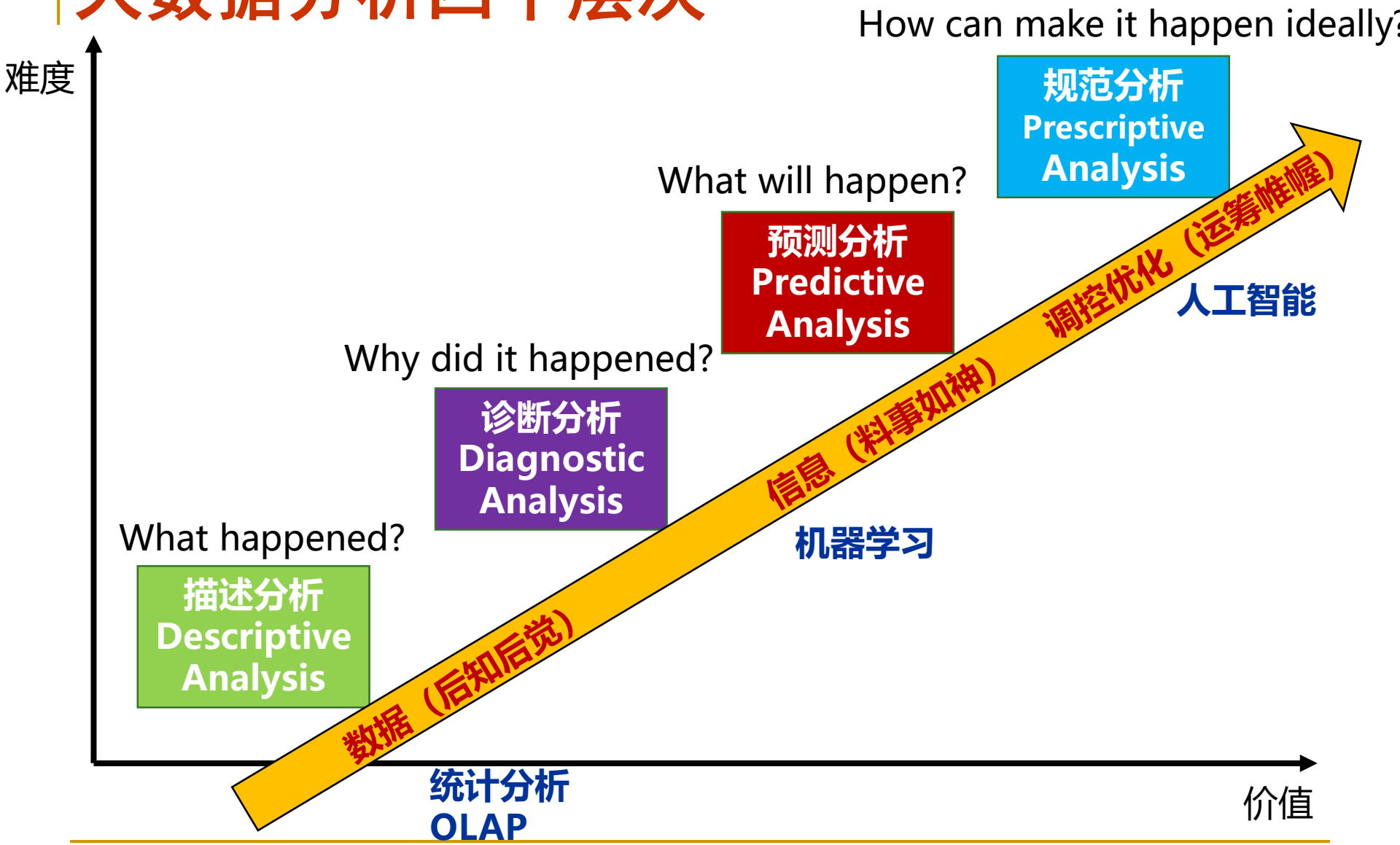
数据体量锐减
价值密度飙升

大数据分析技术

- 大数据分析是大数据价值提炼、实现从数据→知识→决策转换的关键环节，涉及数据科学、统计分析、机器学习与数据智能化应用等多个领域；
- 大数据分析技术范畴：“从数据到信息、从信息到知识、从知识到决策”三个转换过程所涉及的理论、模型、方法与应用技术；



大数据分析四个层次


How can make it happen ideally?
规范分析
Prescriptive
Analysis
What will happen?
预测分析
Predictive
Analysis
Why did it happened?
诊断分析
Diagnostic
Analysis
What happened?
描述分析
Descriptive
Analysis
难度
数据（后知后觉）
信息（料事如神）
调控优化（运筹帷幄）
人工智能
机器学习
统计分析
OLAP
价值

Outline

- 大数据与大数据分析简介
- 大数据分析技术与系统
- 大数据统计分析
- 大数据机器学习
- **数据驱动的自然语言处理**
- 文本大数据分析
- 知识图谱与知识计算
- 大图数据分析
- 社会媒体分析
- 数据与算法安全

内容

- 关键技术
  - 语法分析
    - 词法分析
      - 中文分词
      - 词性标注
    - 句法分析
  - 语义分析
    - 语义表示
  - 内容分析
    - 信息抽取
    - 文本分类
    - 情感分析
  - 机器翻译
  - 问答系统

基于长短时记忆网络的中文分词

- 传统的统计方法严重依赖于特征的设计，而手工提取特征非常地费时费力，并且还存在错误传递问题
- 对一个句子正确地切分，需要能够捕捉一些远距离的信息

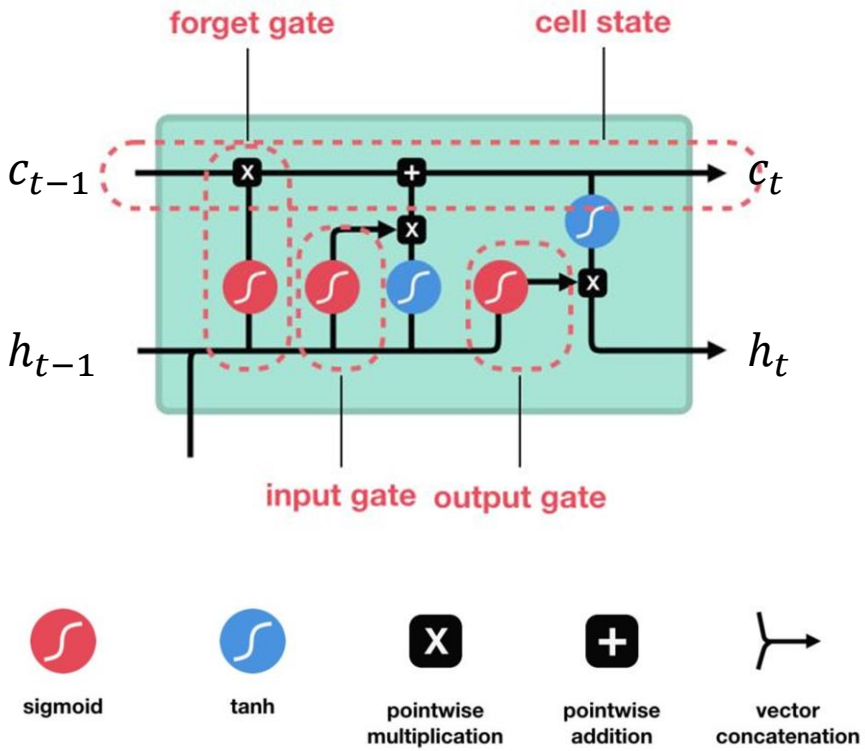
如果不能利用“冬天”和“夏天”，很难对“能穿多少穿多少”进行分词。
LSTM网络可以很好地学习长距离信息。

- Chen等人将长短时记忆网络(Long-Short Term Memory Network, LSTM)用于中文分词

来源：Xinchi Chen et al, Long Short-Term Memory Neural Networks for Chinese Word Segmentation, EMNLP 2015

长短时记忆网络(LSTM)

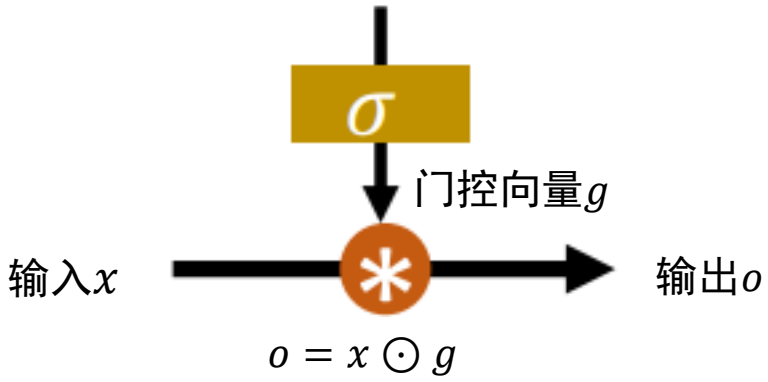
- LSTM的细胞结构
  - 两个状态向量c和h，c是LSTM的内部状态向量，h是LSTM的输出向量
  - 三个门：遗忘门、输入门、输出门

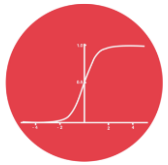


来源：Hochreiter, Sepp, and Jürgen Schmidhuber, Long short-term memory, Neural computation 9.8 (1997)

长短时记忆网络(LSTM)

- Sigmoid门控机制
  - 本质上是一种控制数据流通的手段
  - 可类比于水阀门：当水阀门全部打开时，水流畅通无阻地通过；当水阀门全部关闭时，水流完全被隔断
  - 在LSTM中，阀门开放的程度利用门控向量g表示，具体通过$\sigma()$激活函数将其限缩在[0,1]区间
    - 当g向量某一位为0时，门控全部关闭，输出o在该位上的值为0
    - 当g向量某一位为1时，门控全部打开，输出o在该位上的值与输入x相同





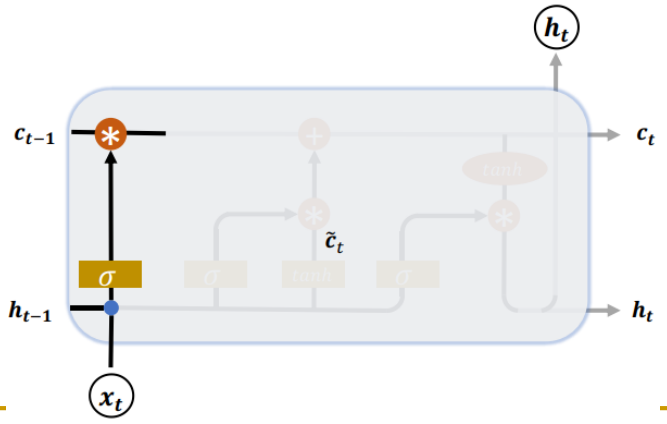
sigmoid将输入压缩到0-1之间

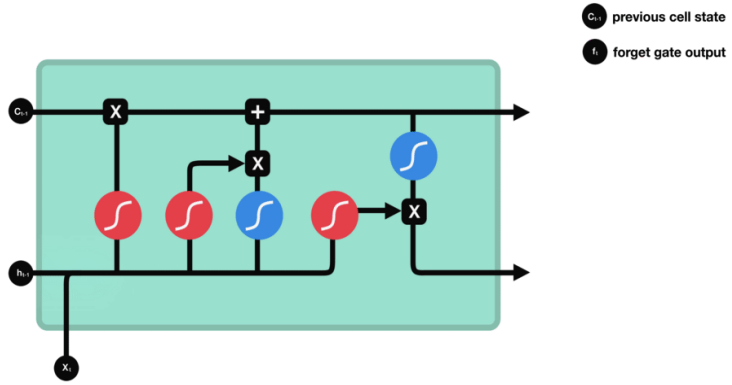
长短时记忆网络(LSTM)

- 遗忘门
  - 遗忘门的功能是决定应丢弃或保留哪些信息。作用于LSTM的状态向量c上，用于控制上一时间戳的记忆c_{t-1}对当前时间戳的影响
  - 具体将前一隐藏状态的信息h_{t-1}和当前输入信息x_t同时传递到sigmoid函数中去，得到遗忘门门控向量：

$$g_f = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

  - 经过遗忘门后，LSTM的状态向量变为：$g_f \odot c_{t-1}$





长短时记忆网络(LSTM)

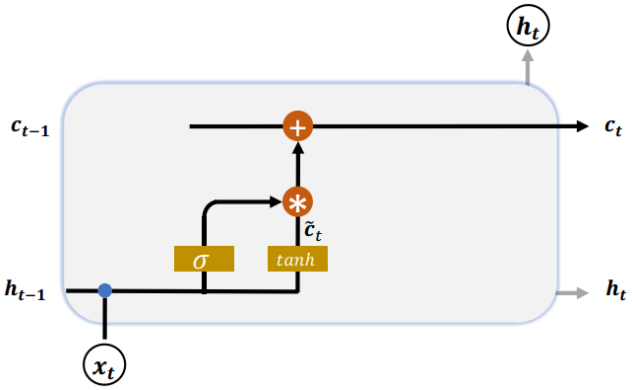
- 输入门
  - 用于更新细胞状态，亦可认为是控制LSTM对输入的接收程度
  - 首先，通过对将前一隐藏状态的信息h_{t-1}和当前输入信息x_t做非线性tanh变换，得到新的输入$\tilde{c}_t$量

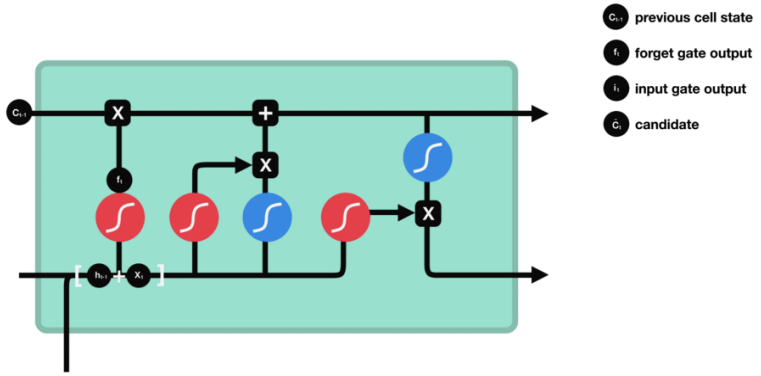
$$\tilde{c}_t = tanh(W_c[h_{t-1}, x_t] + b_c)$$

  - 其次，将前一隐藏状态的信息h_{t-1}和当前输入信息x_t输入sigmoid得到输入门门控向量：

$$g_i = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

  - 经过输入门后，待写入记忆的向量为：$g_i \odot \tilde{c}_t$

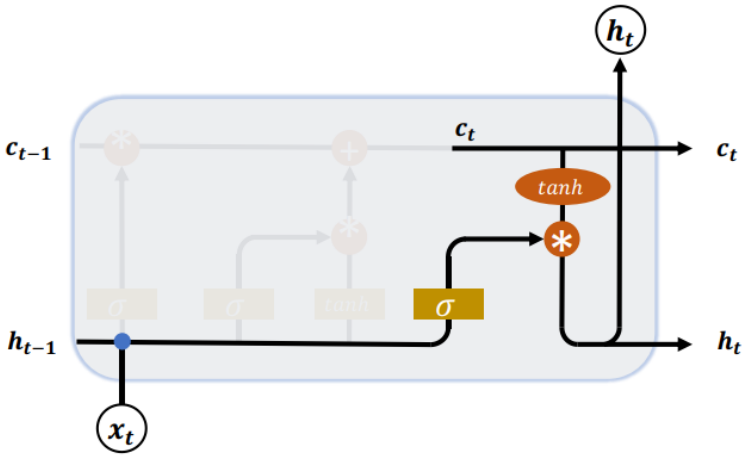


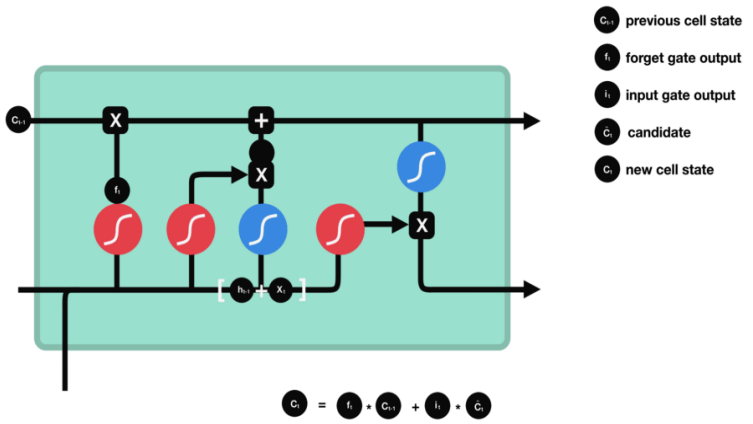


长短时记忆网络(LSTM)

- 更新细胞状态
  - 在遗忘门和输入门的控制下，LSTM接收来自二者的新输入，逐点相加后更新得到当前时间戳的状态向量c_t：

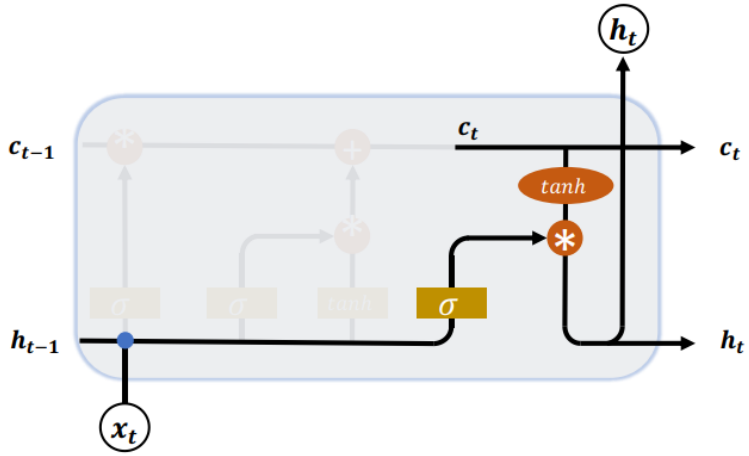
$$c_t = g_i \odot \tilde{c}_t + g_f \odot c_{t-1}$$

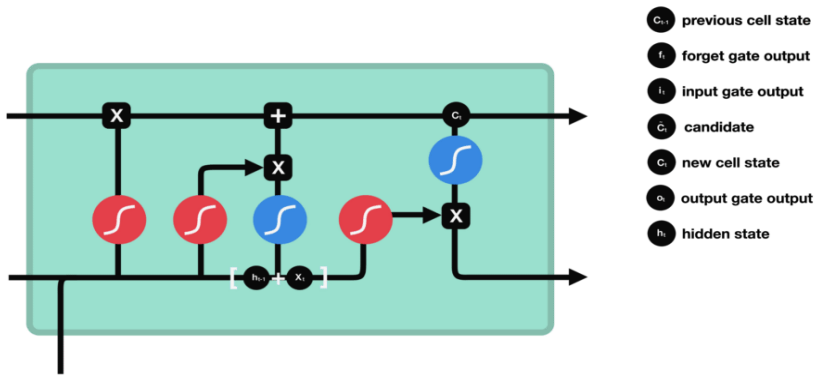




长短时记忆网络(LSTM)

- 输出门
  - 输出门用来确定下个隐藏状态的值，它包含了先前输入的信息
  - 首先，将前一隐藏状态和当前输入传递到sigmoid函数中，得到输出门门控向量：$g_o = \sigma(W_o[h_{t-1}, x_t] + b_o)$
  - 然后，将新得到的细胞状态传递给tanh函数：$\tanh(c_t)$
  - 最后，将tanh的输出与输出门门控向量逐点相乘，确定隐藏状态应携带的信息：$h_t = g_o \odot \tanh(c_t)$

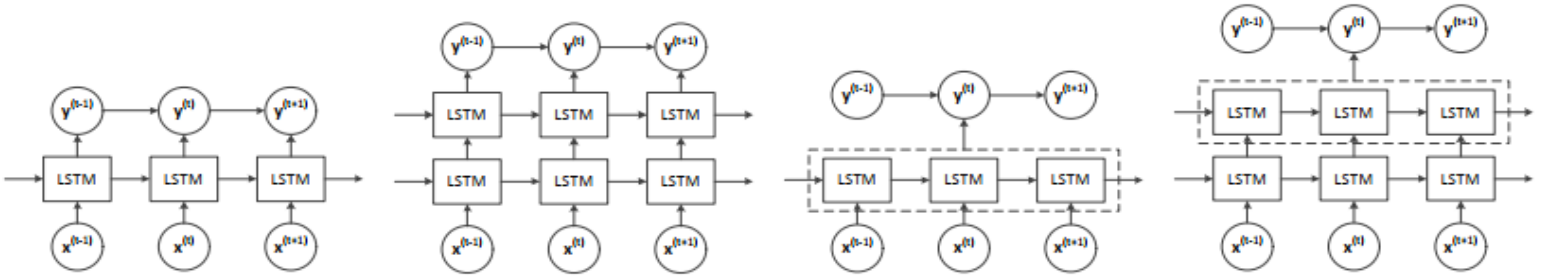




  - 将隐藏状态h_t作为当前细胞的输出，把新的细胞状态c_t和新的隐藏状态h_t传递到下一个时间步长中去

基于长短时记忆网络的中文分词

- LSTM分析模型结构
  - 该模型首先将输入的句子中的每个字都映射成向量表示
  - 然后分别使用四种不同的LSTM+移动窗口的结构提取特征
    - 只使用一层LSTM
    - 只使用双层LSTM
    - 使用一层LSTM，并且联接窗口内LSTM单元输出
    - 使用双层LSTM，并且联接窗口内顶层LSTM单元输出



(a) LSTM-1　(b) LSTM-2　(c) LSTM-3　(d) LSTM-4

  - 通过标签推理层(LSTM层的输出H作为输入)输出每个字对应的标签
    - 进行推断时，输出得分最高的一组句子标注

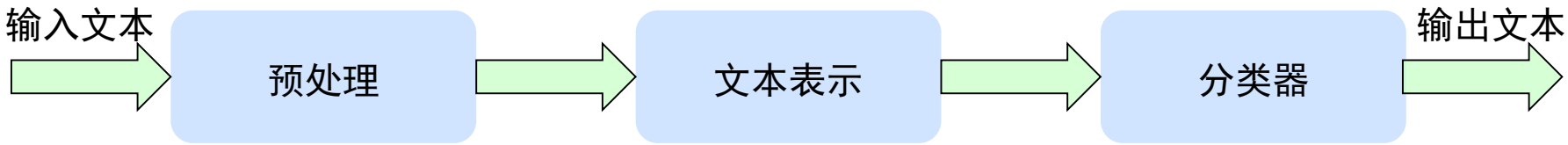
$$s(c^{(1:n)}, y^{(1:n)}, \theta) = \sum_{t=1}^{n} \left(\mathbf{A}_{y^{(t-1)} y^{(t)}} + \mathbf{y}^{(t)}_{y^{(t)}} \right)$$

文本分类

- 文本分类是在自然语言理解中一项基础的任务，是指将一段文本划分到预定义的标签类别中

- 文本分类的效果会直接影响一些下游任务，如：
  - 话题分析
  - 问答系统
  - 自然语言推理

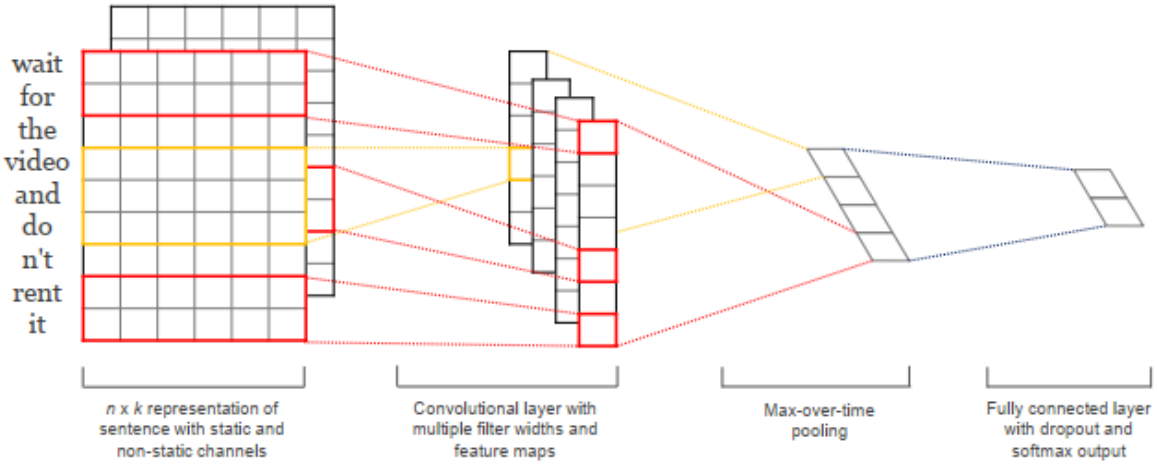
- 文本分类系统



TextCNN模型

- 动机
  - 在由单词的分布式表示组合成的句子表示上，卷积神经网络可以利用卷积核捕捉局部特征
  - 基于卷积神经网络的模型在其他一些传统自然语言处理任务上都取得了不错的效果

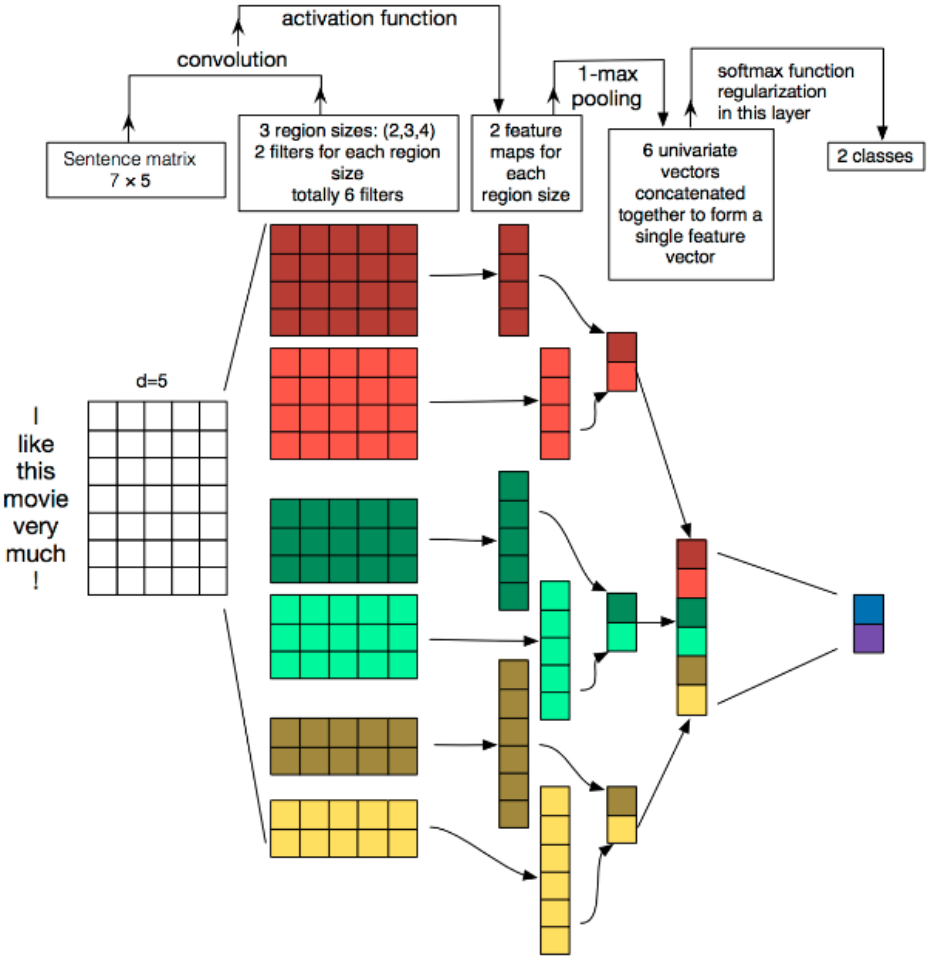
- 模型结构
  - 预训练词向量（Word2Vec, Glove）
  - 卷积层
  - 池化层
  - 全连接层
  - Dropout
  - Softmax



来源：Yoon Kim, Convolutional Neural Networks for Sentence Classification, EMNLP 2014

TextCNN模型

- Embedding：第一层是图中最左边的7乘5的句子矩阵，每行是词向量，维度=5，这个可以类比为图像中的原始像素点
- Convolution：然后经过 filter_sizes=(2,3,4) 的卷积层，每个filter_size有两个输出 channel
- MaxPolling：第三层是一个1-max pooling层，这样不同长度句子经过pooling层之后都能变成定长的表示
- FullConnection and Softmax：最后接一层全连接的softmax层，输出每个类别的概率



来源：Yoon Kim, Convolutional Neural Networks for Sentence Classification, EMNLP 2014

Outline

- 大数据与大数据分析简介
- 大数据分析技术与系统
- 大数据统计分析
- 大数据机器学习
- 数据驱动的自然语言处理
- **文本大数据分析**
- 知识图谱与知识计算
- 大图数据分析
- 社会媒体分析
- 数据与算法安全

内容

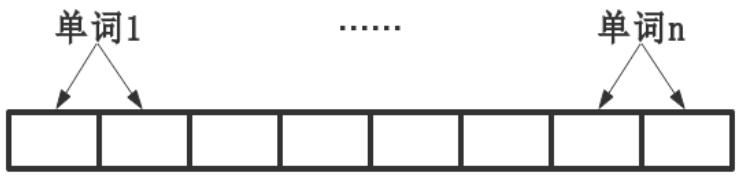
- 文本表达
  - 单词表达方法、句子表达方法
- 文本匹配
  - 基于规则的文本匹配、基于学习的文本匹配
- 文本生成
  - 文本生成任务、方法与评价方式

单词的表示方法

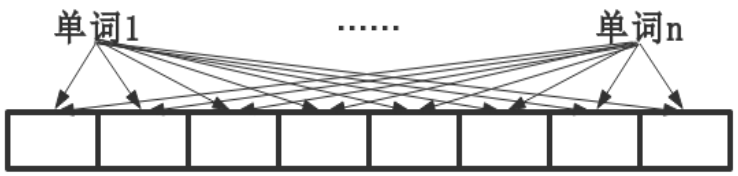
- 局部性表示
  - 独热表示
- 分布式表示
  - 横向组合关系
    - 隐性语义索引(Latent Semantic Indexing, LSI)
    - 概率隐性语义索引(Probabilistic Latent Semantic Indexing, PLSI)
    - 隐性狄利克雷分析(Latent Dirichlet Allocation, LDA)
  - 纵向聚合关系
    - 神经网络概率语言模型(Neural Prob. Language Model, NPLM)
    - 排序学习模型(C&W)
    - 上下文预测模型(Word2Vec)
    - 全局上下文模型(GloVe)

局部性表示 vs. 分布式表示

- 局部性表示(Local Representation)：在将单词表示为向量时，每个单词使用向量中独有且相邻的维度。在这种表示下，单词之间是相互独立的
- 分布式表示(Distributed Representation)：将单词映射到特征空间中，每个单词由刻画它的多个特征来高效表示；在形式上使用稠密实数向量（向量多于一个维度非0，通常为低维向量）来表示单词。分布式表示可以编码不同单词之间的语义关联



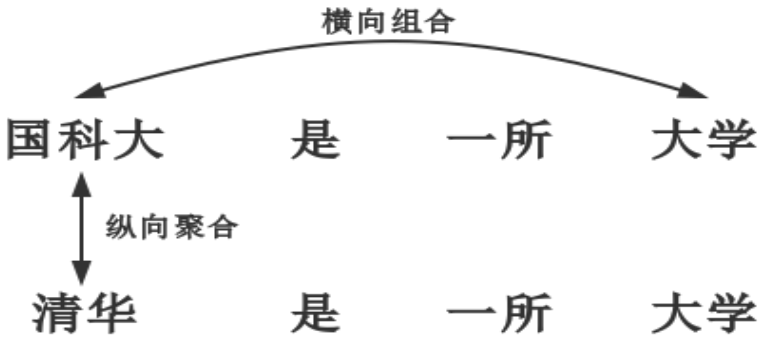
局部性表示



分布式表示

单词的分布式表示

- 分布式表示方法都基于分布语义假设(Distributional Hypothesis)，即单词的语义来自其上下文(context)。因此，
- 所有的分布式表示模型都利用某种上下文的统计信息来学习单词的分布式表示，使用不同的上下文使得模型建模了单词间的不同关系，可分为横向组合关系模型(Syntagmatic Models)和纵向聚合关系模型(Paradigmatic Models)



单词的分布式表示

- 横向组合关系指两个单词可以同时出现在一段文本区域中（如同一个句子），强调它们可以进行组合，在其中往往起到不同的语法作用。如下图中“国科大”和“大学”即存在横向组合关系。对横向组合关系建模的模型通常使用文档作为上下文

- 纵向聚合关系指的是纵向的可替换的关系，强调的是相似的词可以拥有相似的上下文（context）但通常不同时出现。如上图中的“国科大”和“清华”。纵向聚合关系通常使用当前单词周边的单词作为其上下文

单词的分布式表示

- 文档1: I love playing football.
- 文档2: I love playing tennis.
- 文档3: You love playing football.

	$doc1$	$doc2$	$doc3$
I	1	1	0
$love$	1	1	1
$playing$	1	1	1
$football$	1	0	1
$tennis$	0	1	0
you	0	0	1

- 可以观察到：
  - love 和 playing 这两个较强组合关系的词的词表示是相似的；
  - football 和 tennis这两个具有较强替换关系的词的表示是不相似的。

单词的分布式表示

- 文档1: I love playing football.
- 文档2: I love playing tennis.
- 文档3: You love playing football.

	$doc1$	$doc2$	$doc3$
I	1	1	0
$love$	1	1	1
$playing$	1	1	1
$football$	1	0	1
$tennis$	0	1	0
you	0	0	1

- 可以观察到：
  - love 和 playing 这两个较强组合关系的词的词表示是相似的；
  - football 和 tennis这两个具有较强替换关系的词的表示是不相似的。

横向组合关系

- 常用的横向组合关系有:
  - 隐性语义索引(Latent Semantic Indexing, LSI)
  - 概率隐性语义索引(Probabilistic Latent Semantic Indexing, PLSI)
  - 隐性狄利克雷分析(Latent Dirichlet Allocation, LDA)

隐性语义索引(LSI)

- LSI是指通过对词项-文档矩阵C (每行代表一个词项，每列代表一篇文档；元素c_{ij}表示第i个单词在第j篇文档中出现的次数)进行矩阵分解(具体采用SVD分解)来找到它的某个低秩逼近，进而利用得到的低秩逼近形成对词项和文档的新的表示
- 给定$m \times n$的词项-文档矩阵C和正整数k，对C进行LSI的过程如下：
  - 1. 将矩阵C分解为$C = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$；
  - 2. 保持Σ对角线上前k大个奇异值不变，其余元素置为0，得到Σ_k；
  - 3. 计算$C_k = U_{m\times k}\Sigma_{k\times k}V^T_{n\times k}$作为$C$的低秩逼近。一般而言，不同于非负整数构成的较为稀疏的矩阵C，C_k是一个实数构成的稠密矩阵；
    - 3.1. 矩阵$U_{m\times k}$的每一行是相应词项的向量表示，每一维代表该词项在主题空间中的该主题上的映射；
    - 3.2. 矩阵$V_{n\times k}$的每一行是相应文档的向量表示，每一维代表该文档在主题空间中的该主题上的映射。对于给定的文档；

隐性语义索引(LSI)

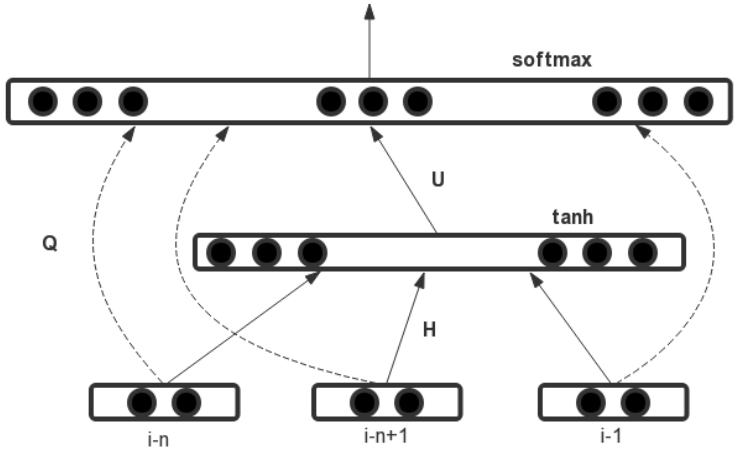
- LSI仅保留了矩阵C中最大的k个奇异值，相当于将原有的词项-文档矩阵从r维降至k维，每一个奇异值可以理解为对应一个“主题”维度，其值的大小表示与这一“主题”的相关程度，因此LSI也是一种主题模型(Topic Model)

- 保持较大的奇异值而将较小的奇异值置为0可以保留文档集中较为重要的信息，并且忽视不重要的细节，从而解决多词一义(synonymy)和语义关联的问题

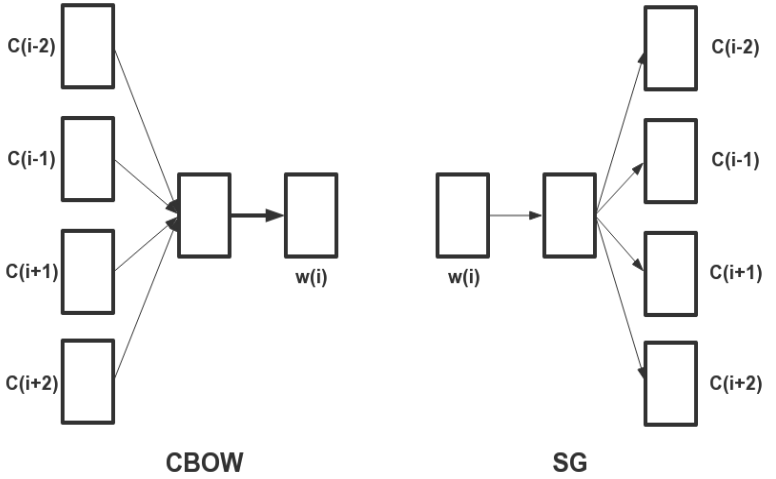
- LSI得到的不是一个概率模型，缺乏统计基础，结果难以直观解释。此外，很难选择合适的k值，而k的选取对结果的影响非常大

纵向聚合关系

- 常用的纵向聚合算法有
  - 神经网络概率语言模型(Neural Probabilistic Language Model, NPLM)
  - 排序学习模型(C&W)
  - 上下文预测模型(Word2Vec)
  - 全局上下文模型(GloVe)等



NPLM模型



CBOW和SG模型框架

排序学习模型(C&W)

- 排序学习模型(C&W) 由Collobert & Weston于2008年提出
- 相比NPLM模型，主要有两点改进：
  - 1. C&W同时使用了单词的上下文，这成为其后学习单词表示的基本做法；
  - 2. C&W对单词序列打分使用了排序损失函数，而不是基于概率的极大似然估计，其损失函数定义为

$$\max[0,1 - s(w,c) + s(w',c)]$$

其中，c 表示单词w 的上下文(context)，w'表示将当前上下文中的单词w替换为一个随机采样出的无关单词w'（负样例）；s为打分函数，分数越高表明该段文本越合理

- 显然，在大多数情况下将普通短语中的特定单词随机替换为其他单词时，得到的都是不正确的短语。因此模型的目标是，尽量使正确短语的得分 比 随机替换后的短语的得分高1

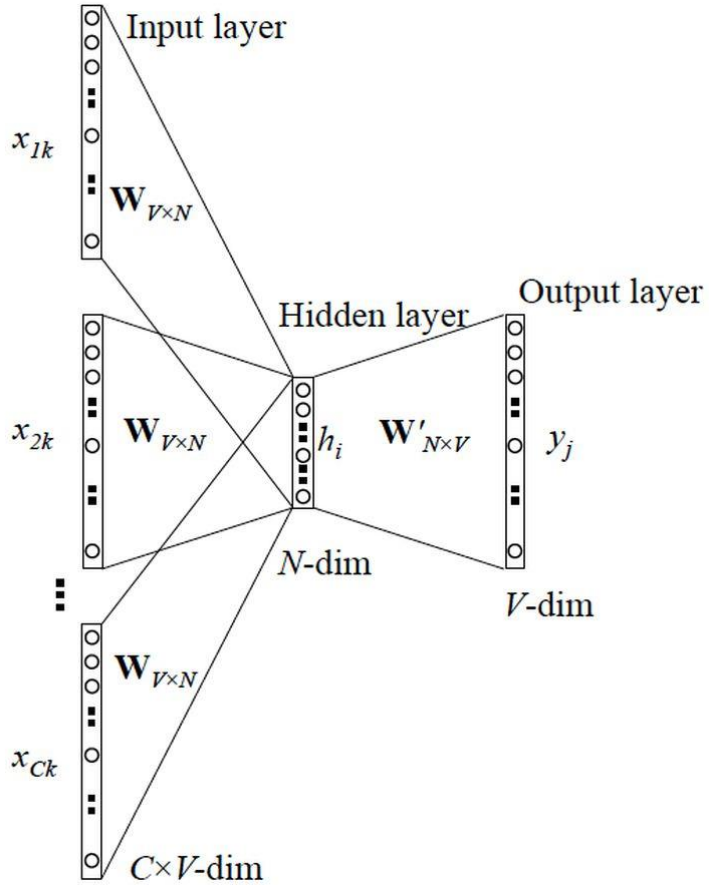
上下文预测模型(Word2Vec)

- 为了更好的利用单词的上下文，Mikolov等人提出了两个简单的神经网络模型CBOW(Continuous Bag of Word)和SG(Skip Gram)进行学习；

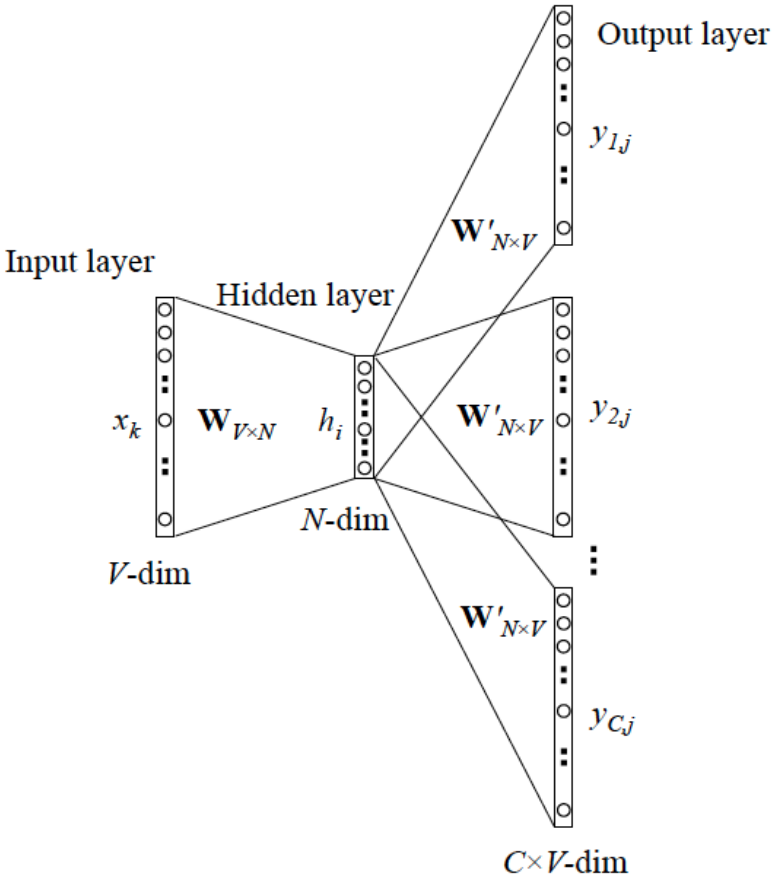
- 相比NPLM模型，CBOW模型去除了中间的非线性隐层，将单词w_i上下文的表示经过求和或平均等计算后，用得到的结果h_i直接预测单词w_i；而SG模型则使用单词w_i预测其上下文中的每一个单词。

- 以短语 “国科大 是 位于 北京 的 大学”为例：
  - CBOW基于上下文“国科大 是 位于 的 大学”预测中心词“北京”
  - SG基于中心词“北京”预测上下文“国科大 是 位于 的 大学”

上下文预测模型(Word2Vec)



CBOW模型



SG模型

全局上下文模型(GloVe)

- 在Word2Vec算法中，主要使用单词的上下文信息获得单词的表示。GloVe模型是一种对词-词共现矩阵进行分解而得到词表示的模型。

	I	$love$	$playing$	$football$	$tennis$	you
I	0	2	0	0	0	0
$love$	2	0	3	0	0	1
$playing$	0	3	0	2	1	0
$football$	0	0	2	0	0	0
$tennis$	0	0	1	0	0	0
you	0	1	0	0	0	0

football和tennis这两个较强替换关系的词的词表示是相似的，而love和playing这两个较强组合关系的词的词表示是不相似的

- 矩阵中的元素值表示的是，以行指标所代表的词作为中心词的窗口内，列指标所代表的词出现的次数，说的简洁一点就是两个词在窗口内的共现次数。
- 上面这个矩阵中，所取的窗口大小为1。比如，以love作为中心词、窗口大小为1的窗口就是“I, love, playing”、“I, love, playing”、“You, love, playing”，那么在窗口内love和playing共现了3次，所以该矩阵的第二行第三列就是3。

全局上下文模型(GloVe)

- 相比Word2Vec算法，GloVe算法利用单词的共现信息，将全文的统计信息与句子的信息相结合，以期得到单词在语义和语句上更好的表达

- 令单词w_i出现的次数为X_i，单词w_i与w_k同时出现的次数为X_{ik}，则在单词w_i出现的情况下单词w_k出现的条件概率为$P(w_k|w_i) = \frac{X_{ik}}{X_i}$

- 研究发现，条件概率的比值$ratio_{i,j,k} = \frac{P(w_k|w_i)}{P(w_k|w_j)}$存在如下规律：

$ratio_{i,j,k}$	单词 j, k 相关	单词 j, k 不相关
单词 i, k 相关	趋于1	很大
单词 i, k 不相关	很小	趋于1

全局上下文模型(GloVe)

- 基于这一观察，对每个单词对相应的向量定义如下的软约束

$$v_i^T v_j + b_i + b_j = log\, X_{ij}$$

其中，v_i和v_j是单词w_i和w_j的向量，b_i和b_j为对应w_i和w_j的偏差，X_{ij}为权重项，正比于单词w_i和w_j共现的次数

- 进一步，定义如下的目标函数

$$J = \sum_{i,j=1}^{N} f(X_{ij})(v_i^T v_j + b_i + b_j - log\, X_{ij})^2$$

其中，

$$f(X_{ij}) = \begin{cases} \left(\frac{X_{ij}}{XMAX}\right)^{\alpha} & if\, X_{ij} < XMAX \\ 1 & \text{Otherwise} \end{cases}$$

$f()$防止只从共现率很高的单词对中学习

句子的表示方法

- 传统方法
  - 词集模型
  - 词袋模型
  - TF-IDF表示
- 分布式表示方法
  - 主题模型
  - 基于单词分布式表示组合的表示方法
  - 由原始语料直接学习的表示方法

词袋模型

- 词袋模型(Bag of Words)是在词集模型的基础上，考虑了单词出现的次数，因此，在词袋模型中，句子向量中每个单词对应的位置上记录的是该单词出现的次数，这也体现了各个单词在该句子中的重要程度

- 示例：
  - 句子：“我 来自 中国 科学院 大学，他 在 中国 科学院 计算所 学习”
  - 单词表vocab={我:0， 来自:1， 中国:2， 科学院:3， 大学:4， 他:5， 在:6， 计算所:7， 学习:8}
  - 词袋模型向量表示：(1， 1， 2， 2， 1， 1， 1， 1， 1)

TF-IDF模型

- TF-IDF（Term Frequency－Inverse Document Frequency)是一种用于信息检索与数据挖掘的常用加权技术。TF是词频(Term Frequency)，IDF是逆向文档频率(Inverse Document Frequency)

- TF-IDF的主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为该词或者短语具有很好的类别区分能力，适合用来分类

- TF计算公式如下：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中，$n_{i,j}$是单词t_i在文档d_j中的出现次数，分母是文档d_j中所有单词的出现次数之和。

TF-IDF模型

- IDF是对一个词语普遍重要性的度量。某一特定词语的IDF，可由总文档数目除以包含该词语之文档的数目，再将得到的商取对数得到。具体地，单词t_i的IDF计算如下：

$$idf_i = log \frac{|D|}{|\{j: t_i \in d_j\}| + 1}$$

其中|D|表示语料库中的文档总数

- TF-IDF：

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i$$

可以看出，某一特定文档内的高词语频率，以及该词语在整个文档集合中的低文档频率，可以产生出高权重的TF-IDF。因此，TF-IDF倾向于过滤掉常见的词语，保留重要的词语

- 对文档d_j的词集模型在对应每个单词t_i的维度赋予该单词的$\mathrm{tfidf}_{i,j}$值，就得到该文档的TF-IDF表示

文本中的匹配问题

文本匹配是自然语言理解的一个核心问题，许多文本处理的问题可以抽象成文本匹配的问题

信息检索	查询项 ⟷	文档
问答系统	问题 ⟷	答案
对话问题	前文 ⟷	回复
复述问题	原句 ⟷	改写
机器翻译	中文 ⟷	英文

搜索引擎

Google

bing

Baidu 百度

问答系统

YAHOO! ANSWERS

知乎

Baidu 知道

智能助手

Siri

微软小冰

基于启发式规则的文本匹配

- 启发式规则的文本匹配模型直接建模了两段文本共同出现的词的分布。不难理解，在两段文本中同时出现的词，对于度量文本的匹配程度是至关重要的

- 两个经典模型
  - BM25
  - 查询似然模型（Query Likelihood Model）

基于隐语义表达的文本匹配

- 除了两段文本中共同出现的词对计算匹配度有贡献，那些词与词之间的关系（如近义词、包含关系等）也应该考虑进匹配度的计算，因此提出了隐语义表达的文本匹配模型

- 这类方法将一段文本映射到一个向量（离散稀疏向量或者连续稠密向量），然后通过计算向量的相似度来表示文本的匹配度。文档表示的各类方法（如TF-IDF和BM25）可以参考上一节的相关内容

文本匹配的评价方法

- 分类准确率（Accuracy）：用于评价分类任务的指标，对于文本匹配任务，只有两类标签，匹配为1，不匹配为0。因此可以把文本匹配看作是一个二分类问题。使用分类准确率可以方便的评价模型对每一对文本的分类是否正确。分类正确的数量占总测试样本数量的比例就是分类准确率

- P@k（Precision at k）：表示前k个文档的排序准确率。假定预测结果排序后，前k个文档中相关文档的数量为Y_k，那么$P@k$可定义为：

$$P@k = \frac{Y_k}{k}$$

- R@k（Recall at k）：表示前k个文档的排序召回率。假设所有相关文档的总数为N。按照预测结果排序后，前k个文档中相关文档的数量为G_k，那么$R@k$可定义为：

$$R@k = \frac{G_k}{N}$$

文本匹配的评价方法

- MAP（Mean Average Precision）：该指标综合考虑了所有相关文档的排序状况。将所有相关文档在预测结果排序中的位置定义为$r_1, r_2, \cdots, r_G$，则平均精度均值指标可定义为：

$$\mathrm{MAP} = \frac{\sum_{i=1}^{G} P@r_i}{G}$$

- MRR（Mean Reciprocal Rank）：如果只考虑预测结果排序中第一个出现的相关文档的位置r_1，可以定义MRR指标为：

$$\mathrm{MRR} = \mathrm{P}@r_1 = \frac{1}{r_1}$$

文本匹配的评价方法

- nDCG（normalized Discounted Cumulative Gain）归一化折扣累计收益
  - 有些任务当中标注的相关度本身就有大小之分而不是单纯的匹配和不匹配两个级别，这个时候nDCG这个指标就会更加有效。nDCG让相关度越高的排在越前面
  - 给定按照标注的文档相关度排序后的文档相关度值分别为 $\widehat{rel}_1, \widehat{rel}_2, \cdots, \widehat{rel}_N$，若按照预测结果排序后的文档相关度的值分别为 $rel_1, rel_2, \cdots, rel_N$。那么，nDCG指标的定义如下：

$$IDCG = \widehat{rel}_1 + \sum_{i=2}^{N} \frac{\widehat{rel}_i}{\log_2(i+1)}$$

$$DCG = rel_1 + \sum_{i=2}^{N} \frac{rel_i}{\log_2(i+1)}$$

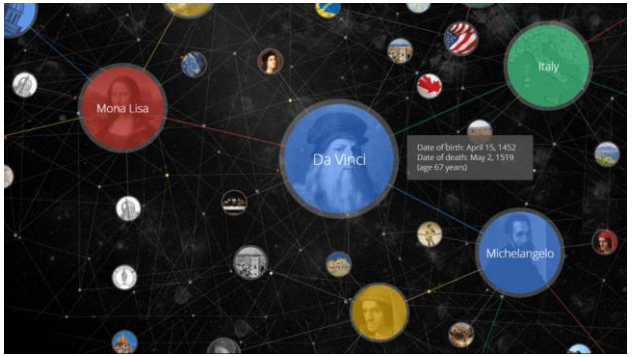
$$nDCG = \frac{DCG}{IDCG}$$

Outline

- 大数据与大数据分析简介
- 大数据分析技术与系统
- 大数据统计分析
- 大数据机器学习
- 数据驱动的自然语言处理
- 文本大数据分析
- **知识图谱与知识计算**
- 大图数据分析
- 社会媒体分析
- 数据与算法安全

知识图谱 (Knowledge Graph)

- 知识图谱本质上是一种语义网络(Semantic Net)，其结点代表实体(entity)或概念(concept)，边代表实体/概念之间的各种语义关系；

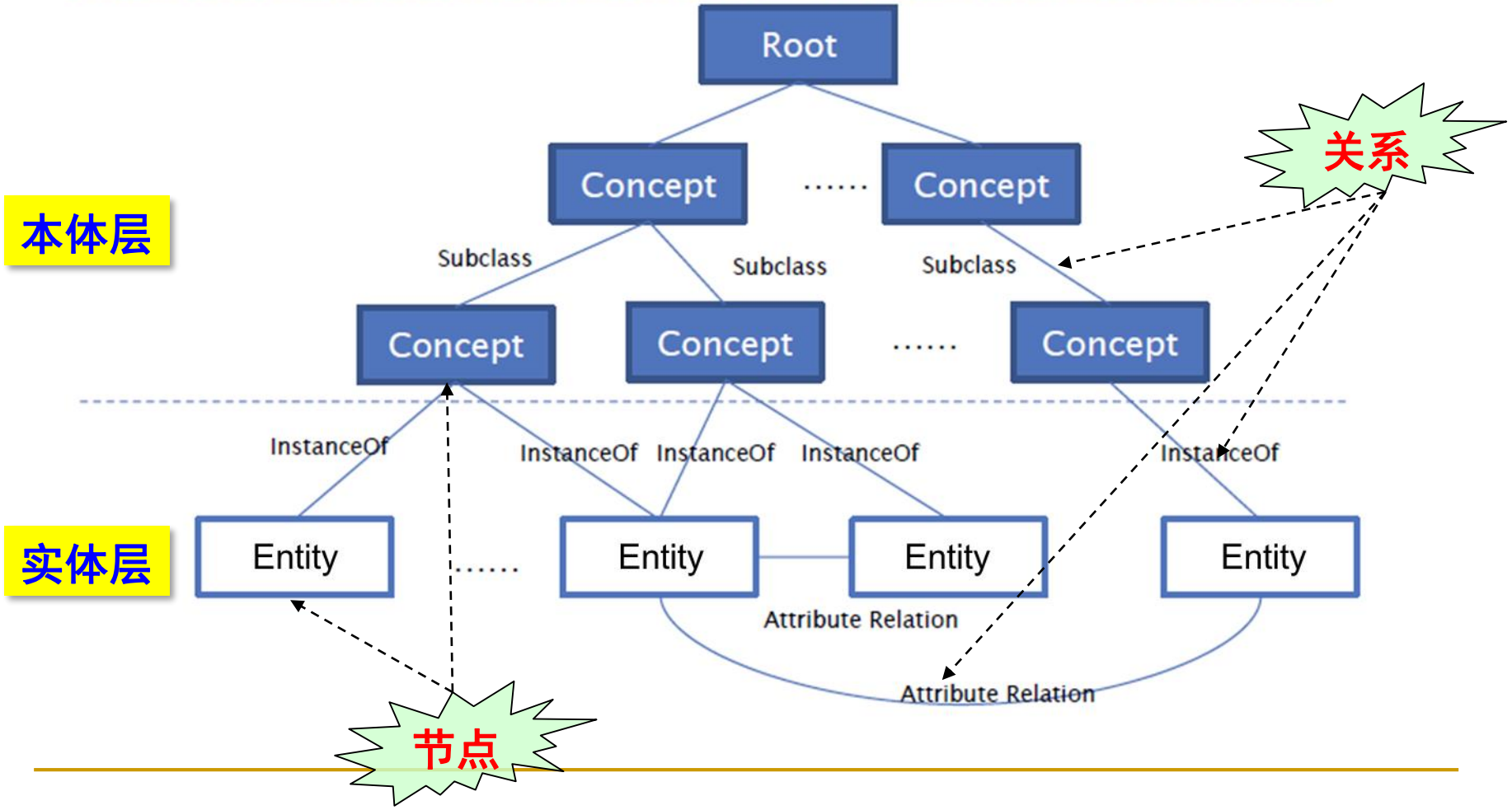


Google, 2012

- A Knowledge Graph (KG) is a system that understands facts about people, places and things and how these entities are all connected;

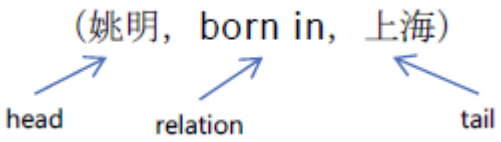
- 知识图谱把不同来源、不同类型的信息连接在一起形成关系网络，提供了从关系的角度去分析问题的能力

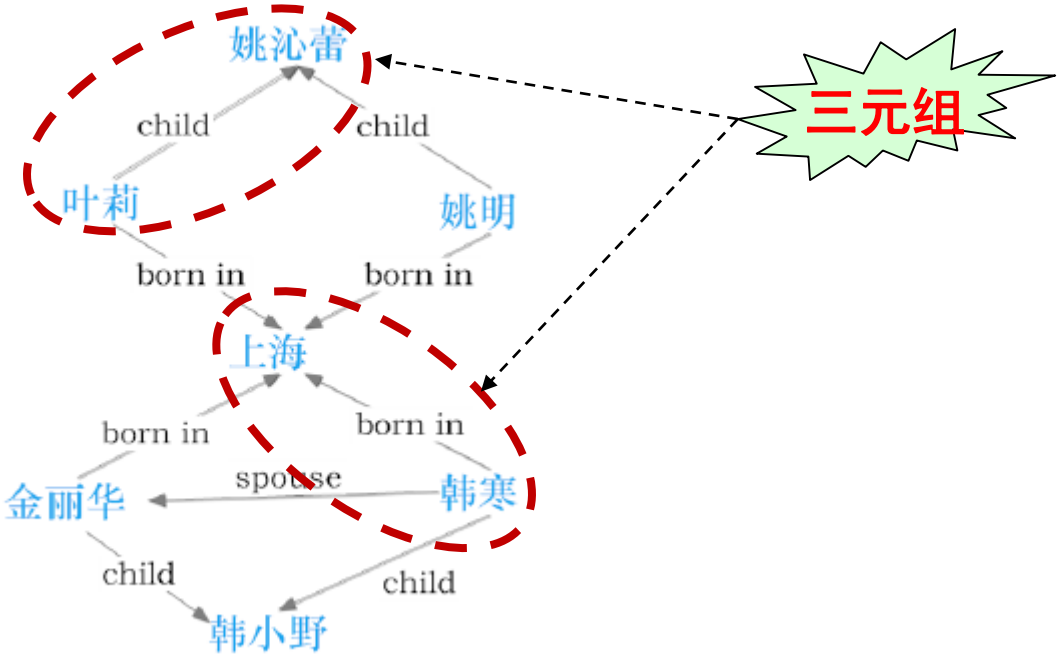
什么是知识图谱?


本体层
实体层
Root
Concept
Concept
Concept
Concept
Concept
Subclass
Subclass
Subclass
InstanceOf
InstanceOf
InstanceOf
InstanceOf
InstanceOf
Entity
Entity
Entity
Entity
Attribute Relation
Attribute Relation
关系
节点

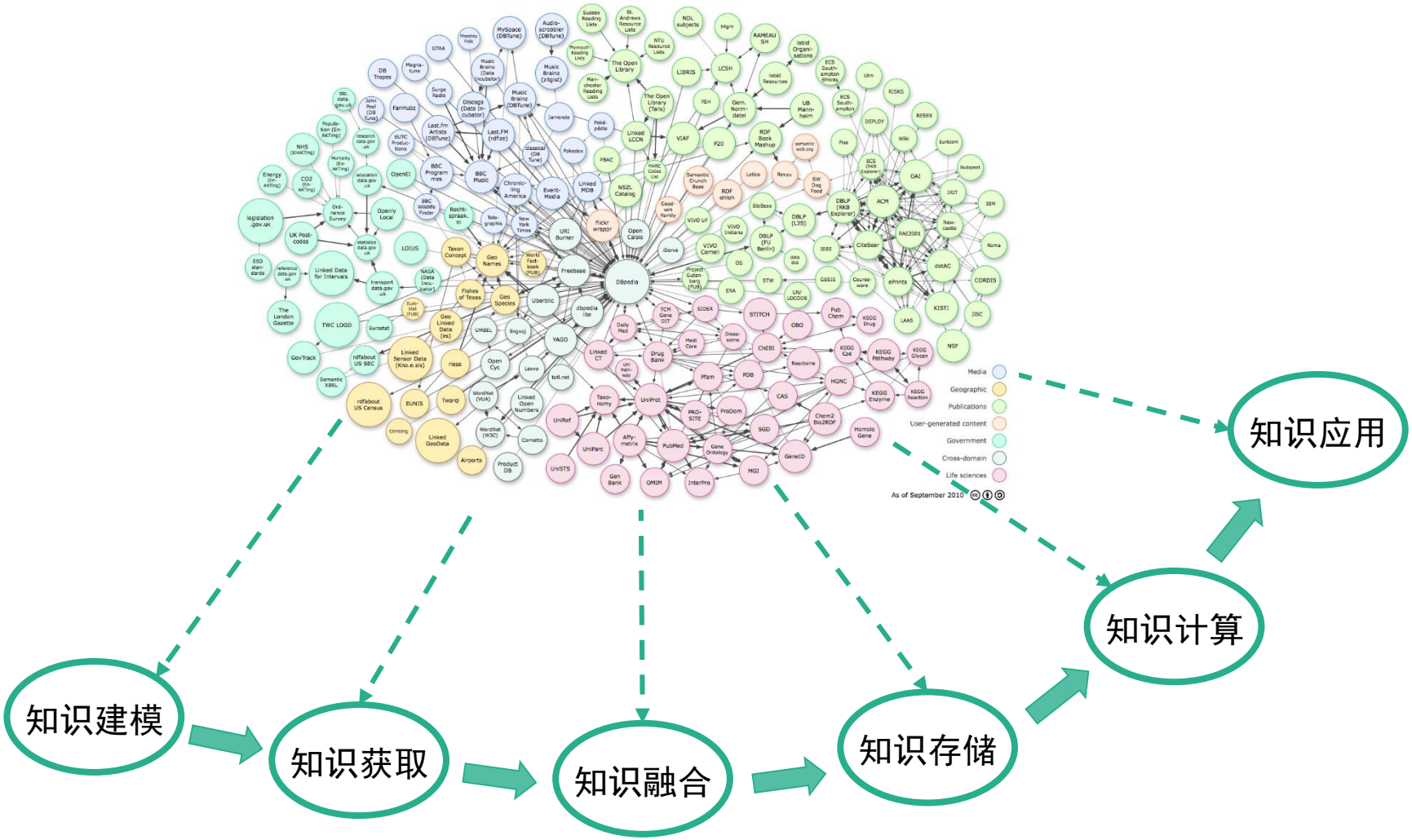
知识图谱中的知识表示：三元组

- 三元组Triple: (head, relation, tail)
  - head：头实体/概念
  - relation：关系/属性
  - tail：尾实体/概念





知识图谱的生命周期



实体抽取

- 实体抽取定义
  - 从原始语料中自动识别出指定类型的命名实体，主要包括实体名（如人名、地名、机构名、国家名等）、缩略词，以及一些数学表达式（如货币值、百分数、时间表达式等）

- 示例

  5月19日下午，史密斯教授做客北京大学海外名师讲堂。

  时间　　人名　　机构名

基于机器学习的方法

- 序列标注
  - 实体标注一般使用BIO模式

(B-begin，I-inside，O-outside)

输入序列	小明	昨天	晚上	在	公园	遇到	了	小红	。
语块	B-NP	B-NP	I-NP	B-PP	B-NP	B-VP		B-NP	
标注序列	B-Agent	B-Time	I-Time	O	B-Location	B-Predicate	O	B-Patient	O
角色	Agent	Time	Time		Location	Predicate	O	Patient	

  - 还有BIOES标注模式

(B-begin，I-inside，O-outside，E-end，S-single)

基于机器学习的方法

- 隐马尔科夫模型
  - 假定分词后的文档词语序列为$W = (w_1, \ldots, w_n)$，$T = (t_1, \ldots, t_n)$为词序列的实体标注结果。模型旨在给定词语序列W的情况下，找出概率最大的标注序列T，即，求使$P(T \mid W)$最大的标注序列

$$T_{max} = arg_T max P(T|W)$$

根据贝叶斯公式，

$$P(T|W) = P(T)P(W|T)/P(W)$$

其中，$P(W)$可以看成一个常数，则有

$$T_{max} = arg_T max P(T)P(W|T)$$

其中，$P(T)P(W|T)$是引入隐马尔科夫模型来计算的参数。如果穷举序列W和T的所有可能情况，这个问题是NP难的

基于机器学习的方法

- 隐马尔科夫模型
  - 按照马尔科夫假设，当前状态t_i只和其前一状态t_{i-1}有关，因此有

$$P(T)P(W|T) \approx \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

其中，$P(w_i|t_i)$表示隐状态为t_i的词语集合中出现w_i的概率，$P(t_i|t_{i-1})$表示上一词语标注为t_{i-1}时，当前词语标注为t_i的转移概率。进一步

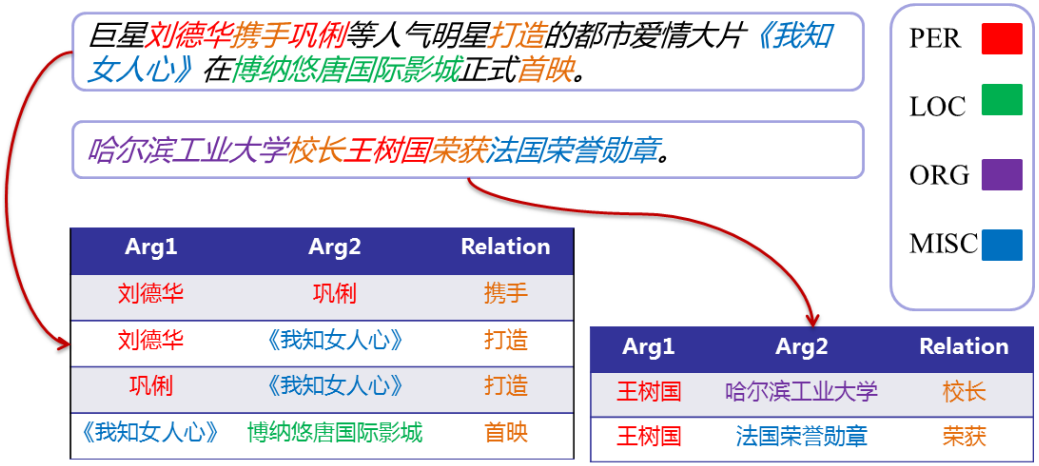
$$T_{max} = arg_T max \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

$$T_{max} = -arg_T min \sum_{i=0}^{n} \{lnP(w_i|t_i) + lnP(t_i|t_{i-1})\}$$

训练时，取$P(w_i|t_i) \approx Count(w_i, t_i)/Count(t_i)$，其中$Count(w_i, t_i)$表示词语$w_i$被标注为$t_i$的次数，$Count(t_i)$表示隐状态$t_i$出现的总次数

关系抽取

- 关系抽取示例



- 抽取方法分类
  - 基于传统机器学习的方法
    - 有监督方法、半监督方法、远程监督方法、无监督方法
  - 基于深度学习的方法
    - 基于经典神经网络的方法、基于BERT的方法

远程监督关系抽取

- 初始动机：通过外部知识库代替人对语料进行标注，从而低成本地获取大量有标注数据 [Mintz et al., 2009]

- 核心思想：如果知识库中存在三元组$\langle e_1, R, e_2 \rangle$，那么语料中所有出现实体对$\langle e_1, e_2 \rangle$的语句，都标注为表达了关系R

- 根据这一假设，对每个三元组$\langle e_1, R, e_2 \rangle$，将所有$\langle e_1, e_2 \rangle$共现的句子都标注标签$R$，用分类方法解决关系抽取问题

远程监督关系抽取

- Riedel等[Riedel et al., 2010]认为Mintz的假设过强，可能引入噪声模式，因而提出“at-least-once”假设：
  - 如果存在三元组$\langle e_1, R, e_2 \rangle$，那么所有$\langle e_1, e_2 \rangle$实体对共现的语句中，至少有一句体现了关系R在这两个实体上成立的事实

- 引入了多实例学习机制，将所有$\langle e_1, e_2 \rangle$共现的句子聚成一个句袋，并将任务由对句子分类变为对句袋分类

基于预训练模型BERT的关系抽取

- 预训练模型是近年来自然语言处理领域取得的非常重要的进展，直接推动了关系抽取研究；

- Wu等人将大规模预训练模型BERT应用于关系抽取，在 SemEval-2010 Task 8数据集上取得SOTA。

来源： Wu, Shanchan, and Yifan He. "Enriching pre-trained language model with entity information for relation classification." Proceedings of the 28th ACM international conference on information and knowledge management. 2019.

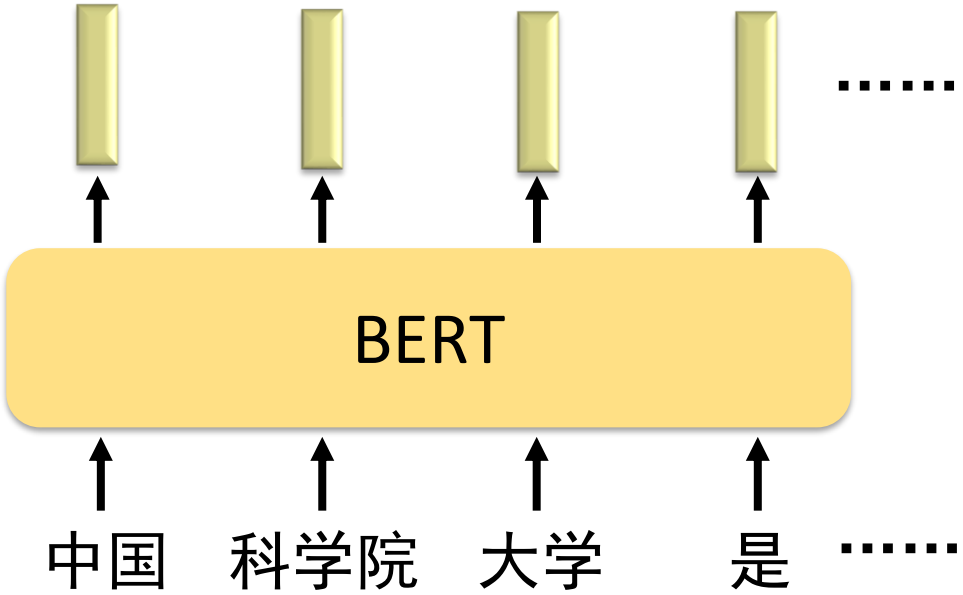
BERT简要介绍

- BERT的全称为基于Transformer的双向编码表征(Bidirectional Encoder Representations from Transformers)，是一个预训练的语言模型；
- BERT没有采用传统的单向语言模型或把两个单向语言模型进行浅层拼接的方法进行预训练，而是采用遮盖语言模型，所以能生成深度的双向语言表征；

- BERT具有两个主要优点
  - 采用遮盖语言模型对双向的Transformer进行预训练，以生成深层的双向语言表征；
  - 预训练后，只需要添加一个额外的输出层进行微调，就可以在各种各样的下游任务中取得最优的表现。在这过程中并不需要对BERT进行任务特定的结构改造；

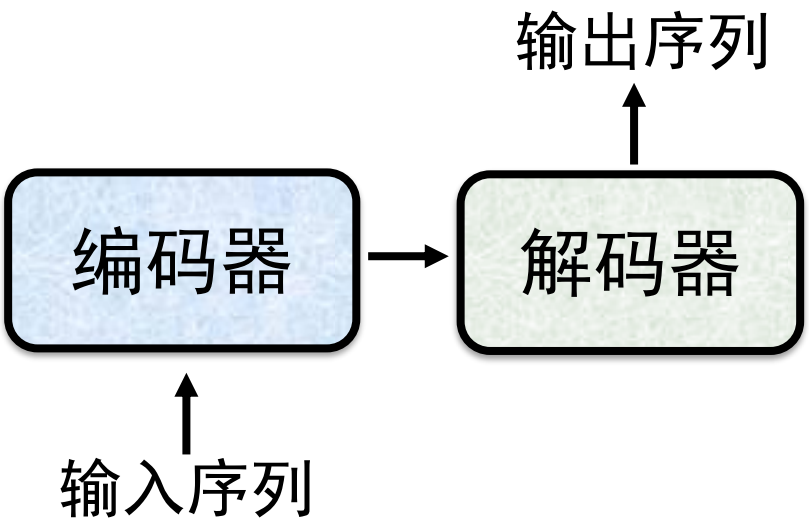
BERT简要介绍

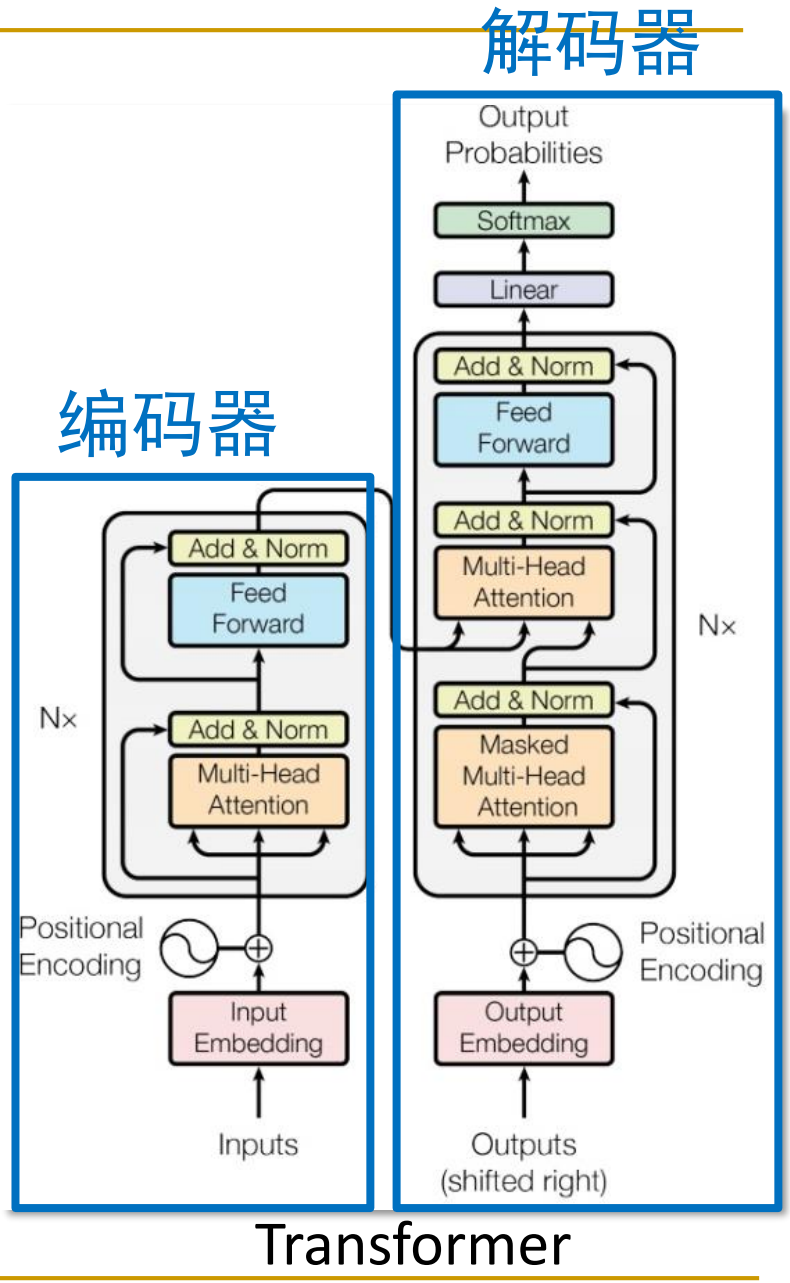
- 输入一个序列，然后输出序列中单词对应的向量表示



Transformer

- Transformer是典型的序列编码模型，分为编码器、解码器两部分；
- 先使用编码器对序列进行编码，再使用解码器解码为所需要的表示；

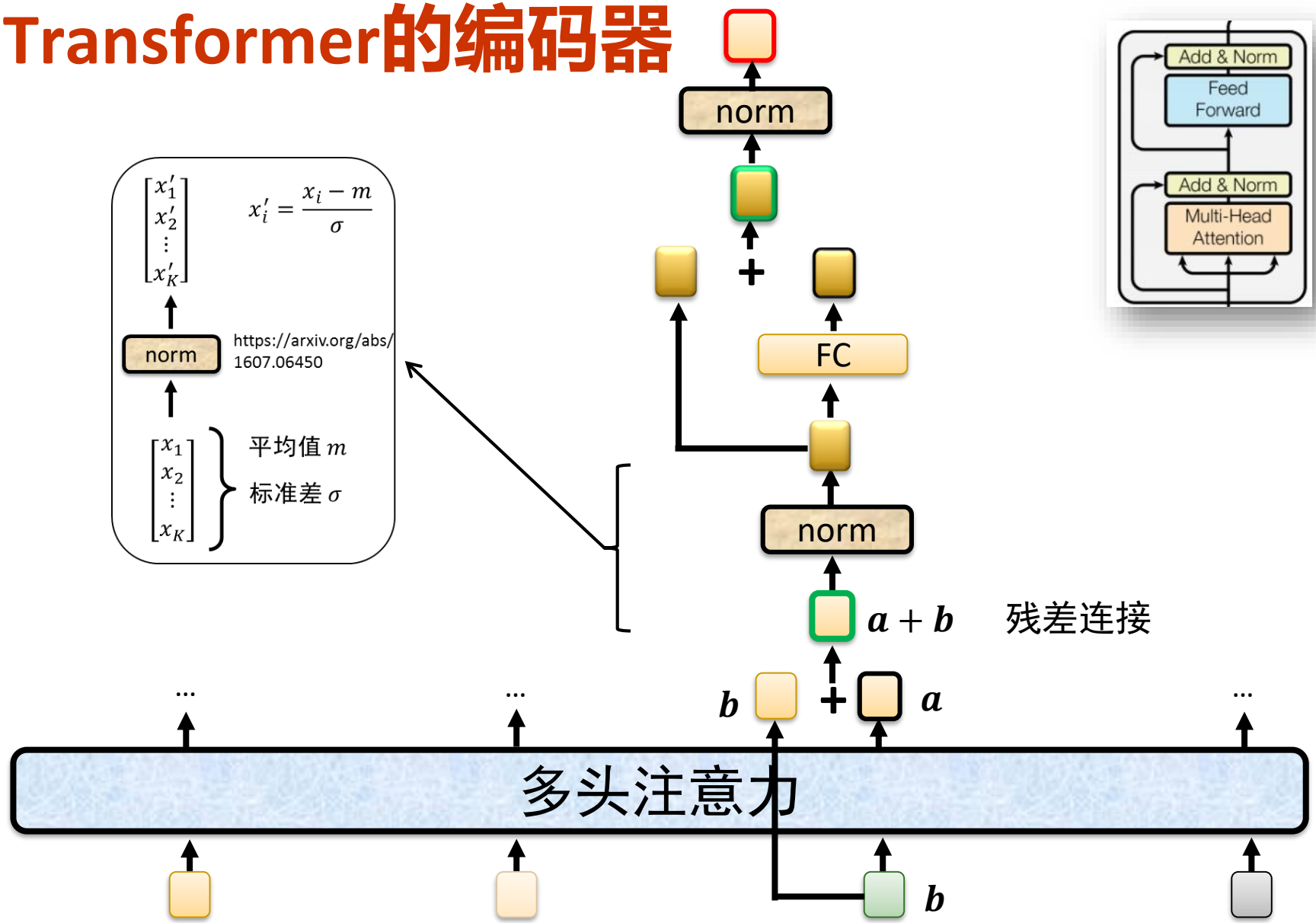




https://arxiv.org/abs/1409.3215

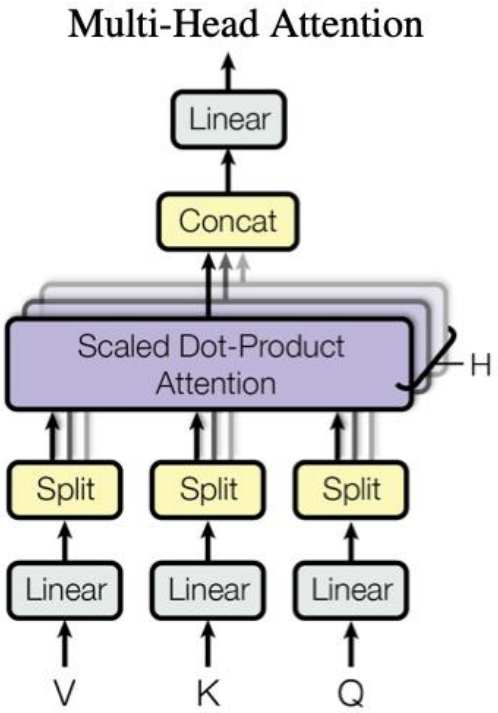
https://arxiv.org/abs/1706.03762

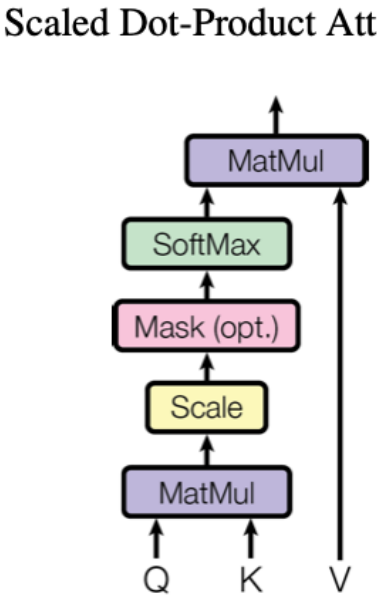
Transformer的编码器



多头注意力

- 多头注意力是由多个缩放点积注意力组成，将每个头的注意力的输出进行拼接得到输出向量；





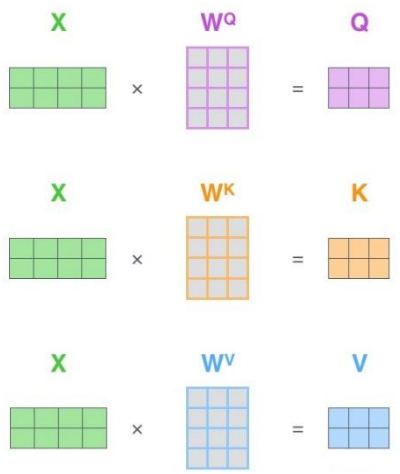
Q表示query，用来匹配K
K表示key，用来被Q匹配
V表示value，要被抽取出来的信息
MatMul表示矩阵乘积

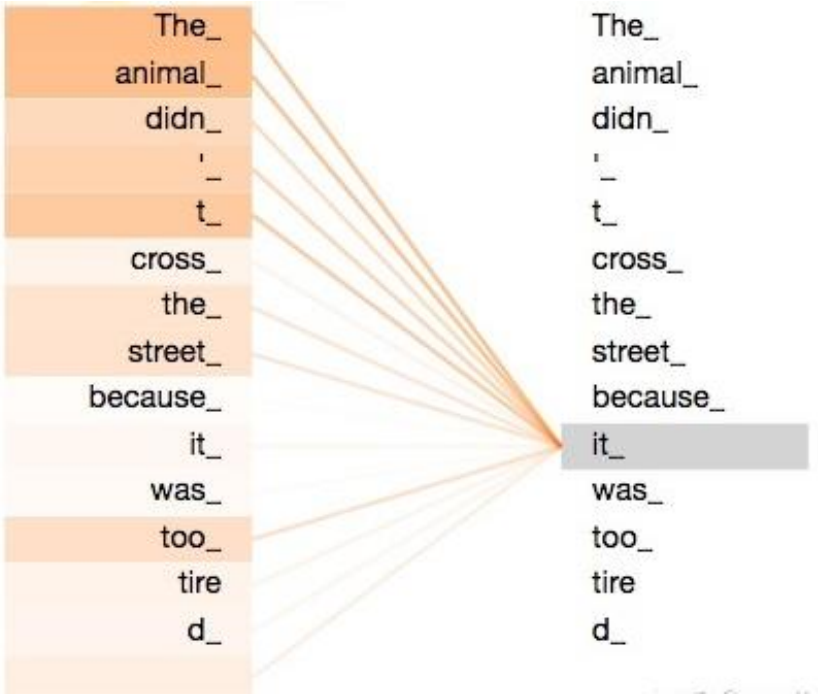
$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,...,\text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Transformer中的自注意力

- Transformer多头注意力中的Q、K、V都是由输入单词投影得到，属于自注意力类型；
- 输出的单词表示由序列中每一个单词表示加权得到；

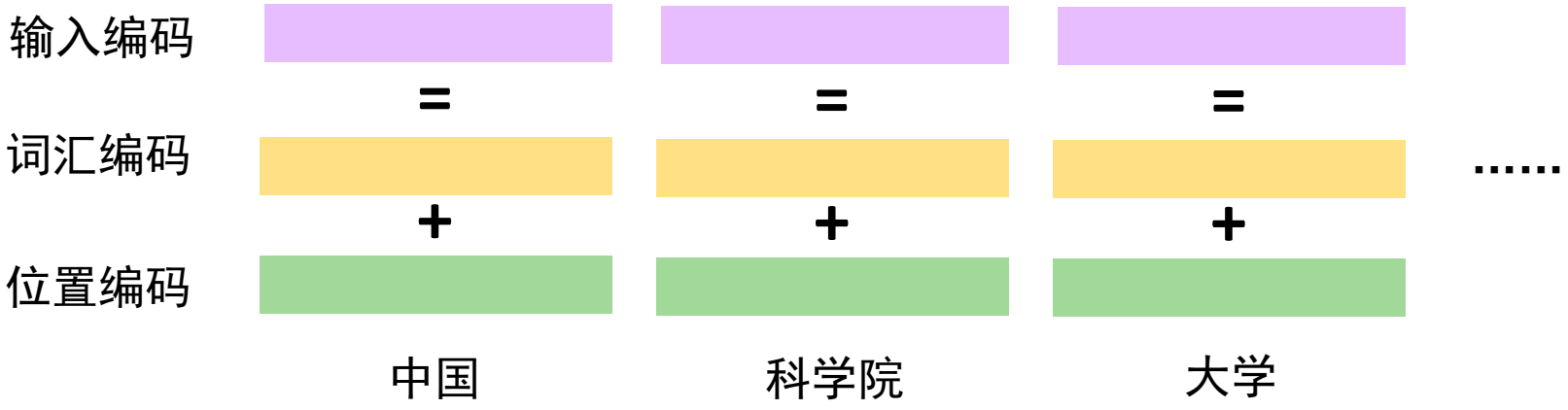




The animal didn't cross the street because it was too tired

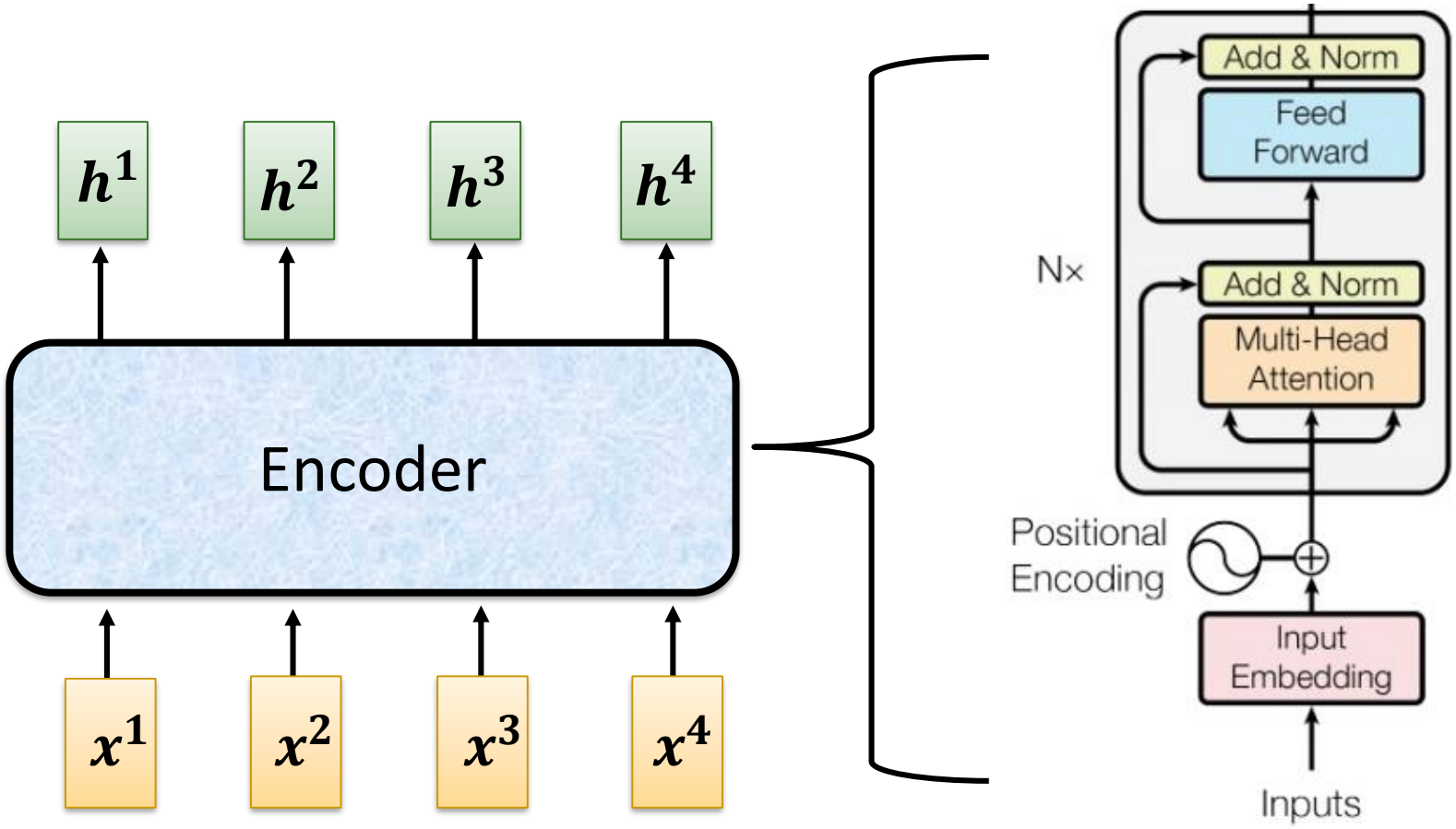
Transformer中的位置编码

- 在词汇编码中没有包含位置信息，然而，词汇的位置信息在序列中往往是关键的；
- 因此，对每个位置初始化一个表示向量，并与单词向量表示相加后作为模型输入；



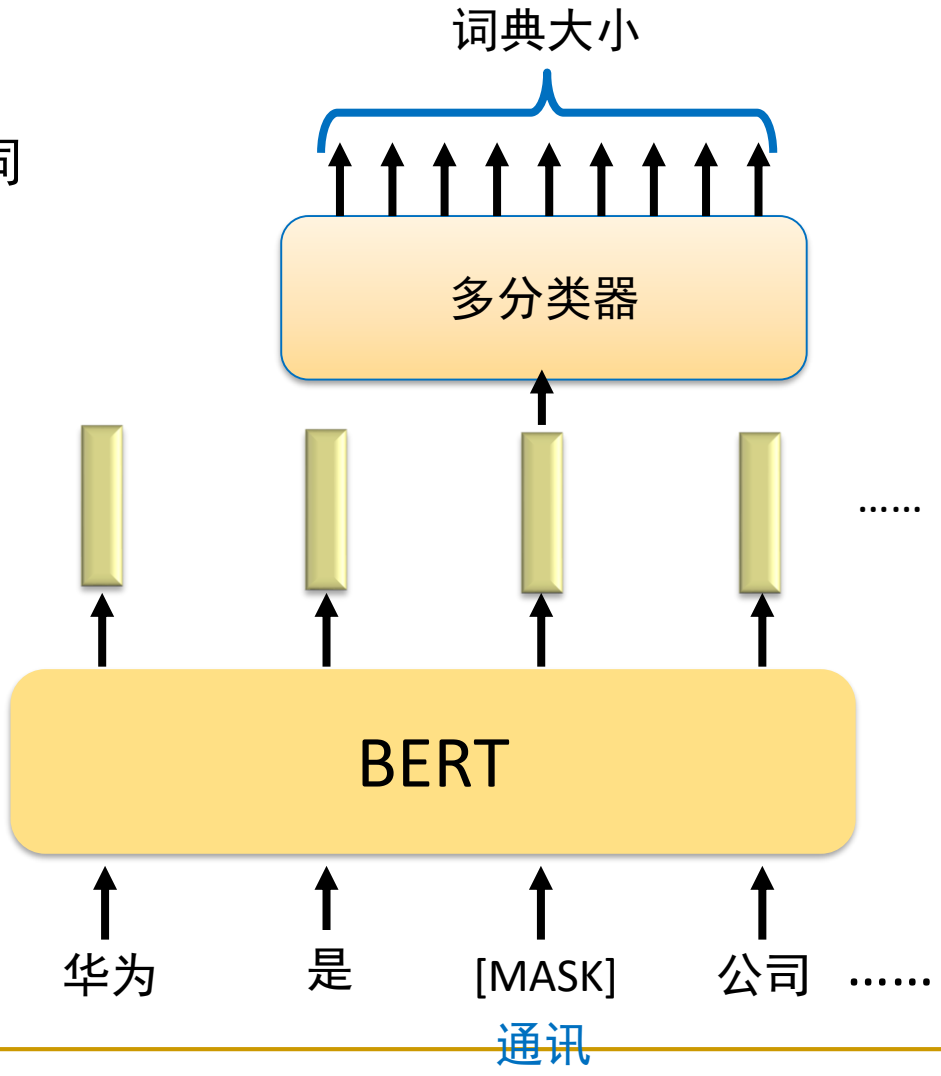
BERT

- BERT是使用多个Transformer编码器堆叠而成



BERT的训练

- 任务1：单词预测
  - 预测被遮盖的单词



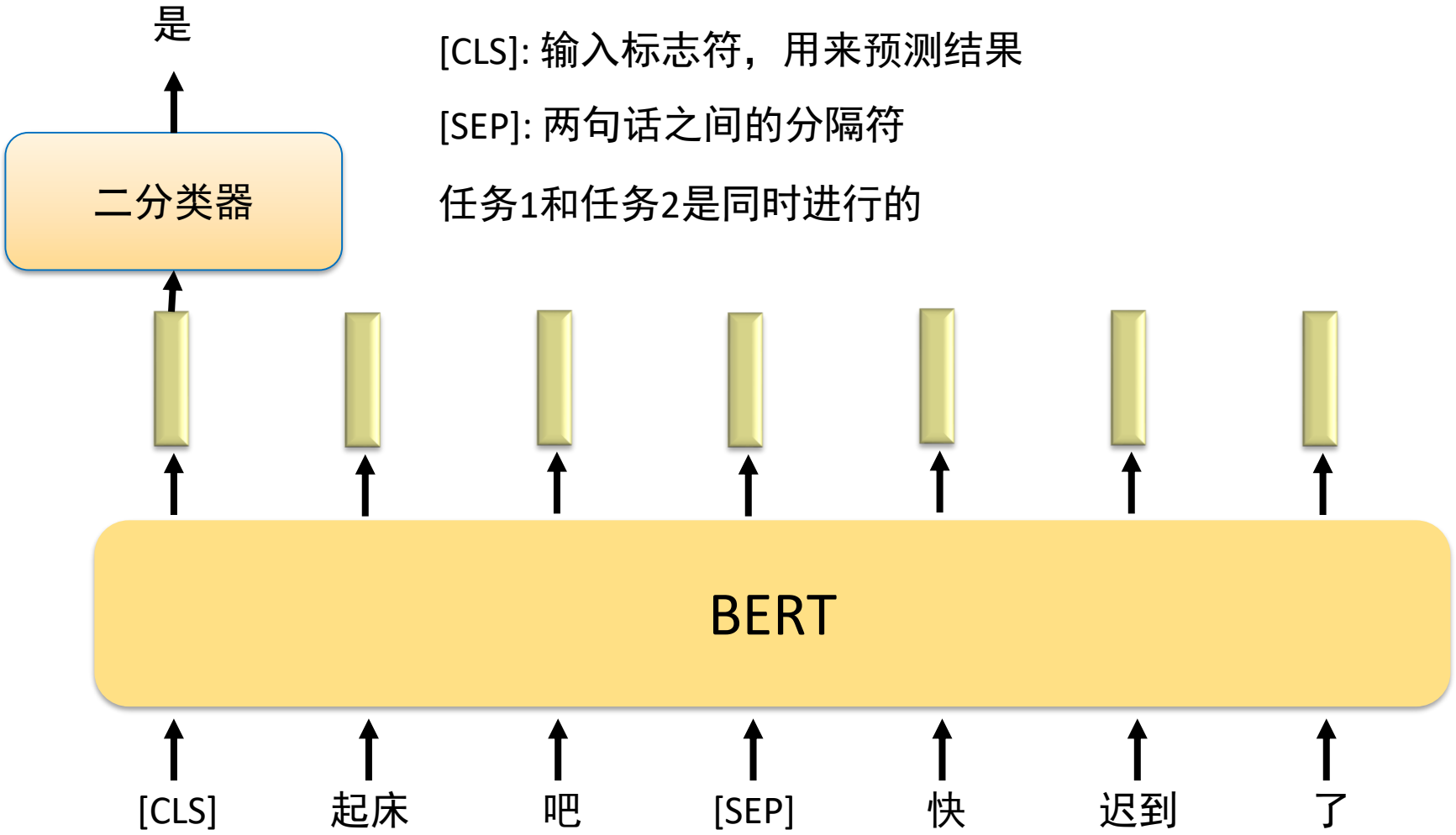
BERT的训练

- 任务 2：下一句话预测

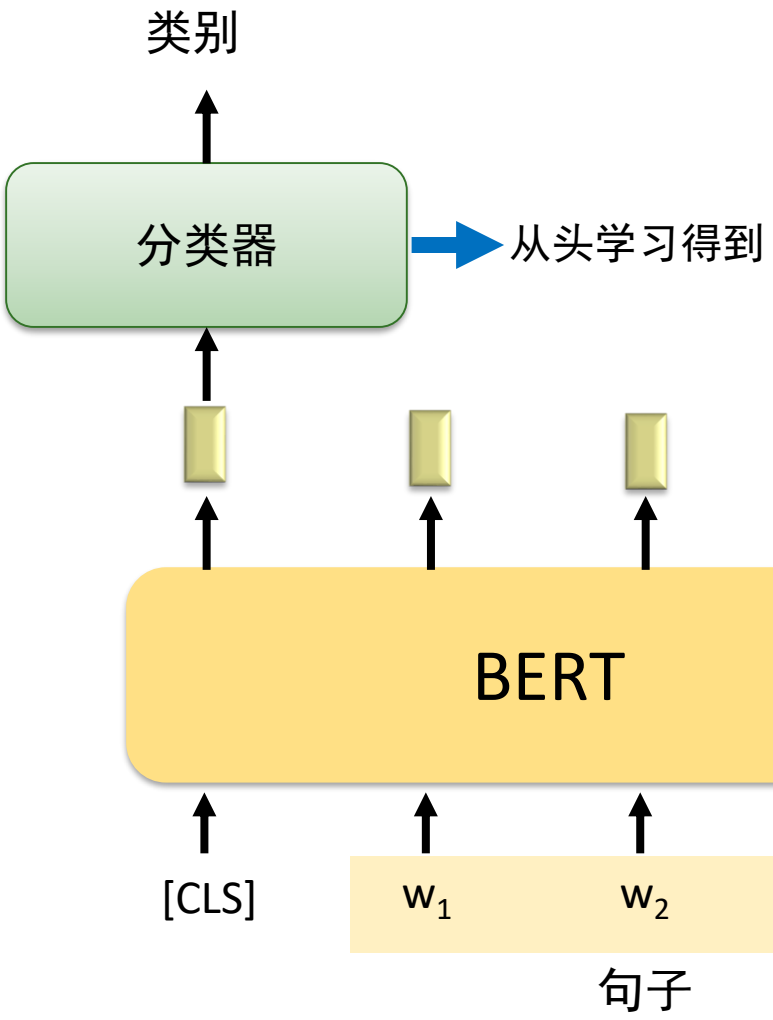
[CLS]: 输入标志符，用来预测结果

[SEP]: 两句话之间的分隔符

任务1和任务2是同时进行的



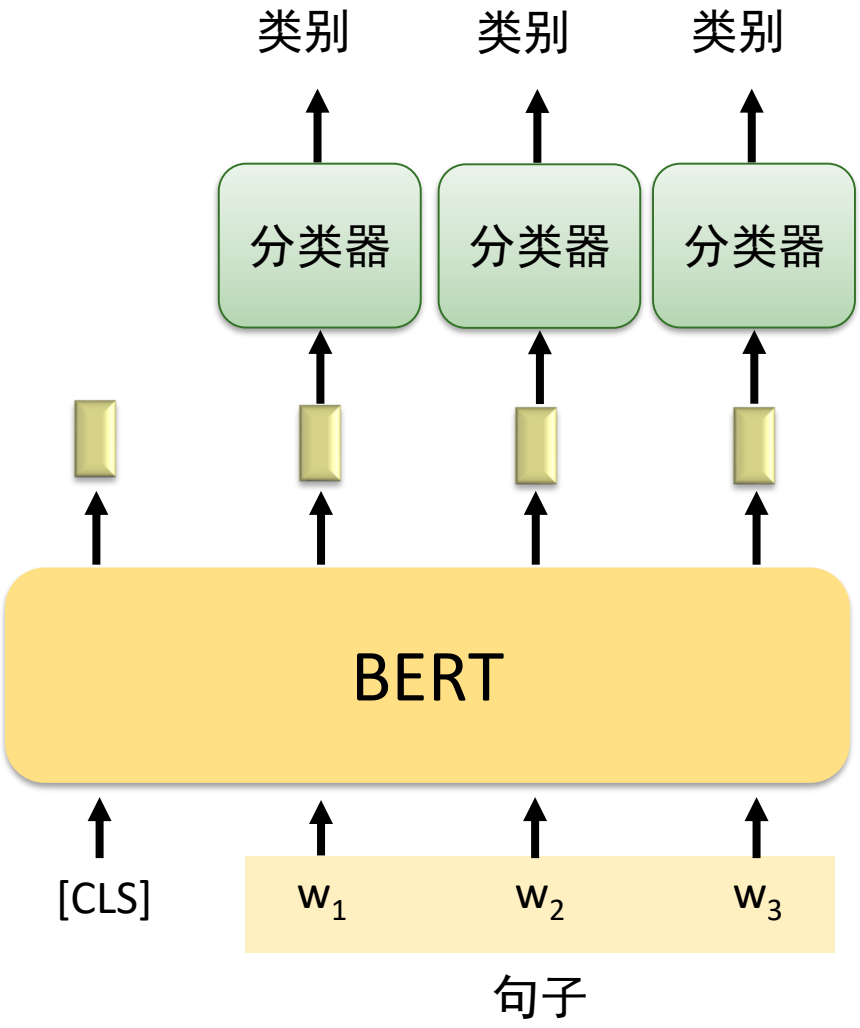
如何使用BERT – 样例1



输入：单个句子，
输出：类别
例子：情感分类，文档分类

读取预训练模型后进行微调

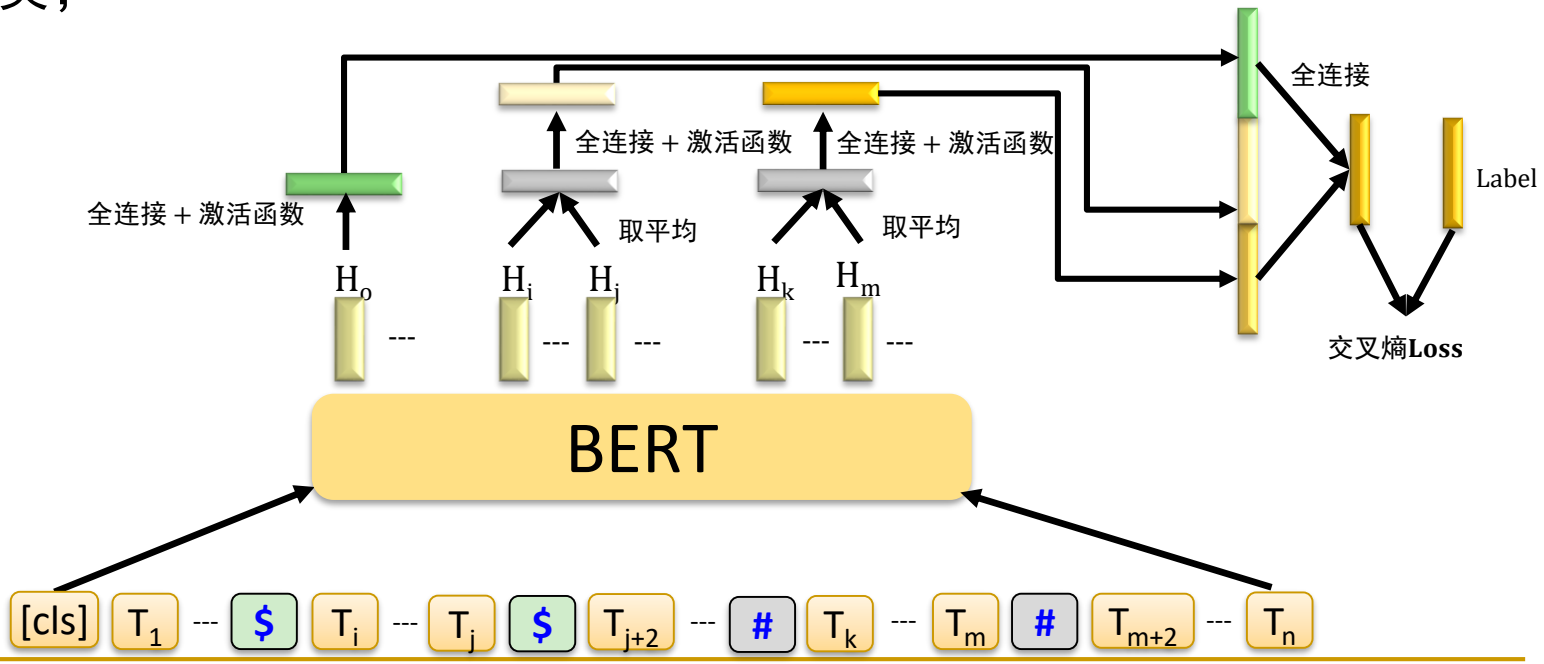
如何使用BERT – 样例2



输入：单个句子，
输出：每一个单词的类别
例子：命名实体识别

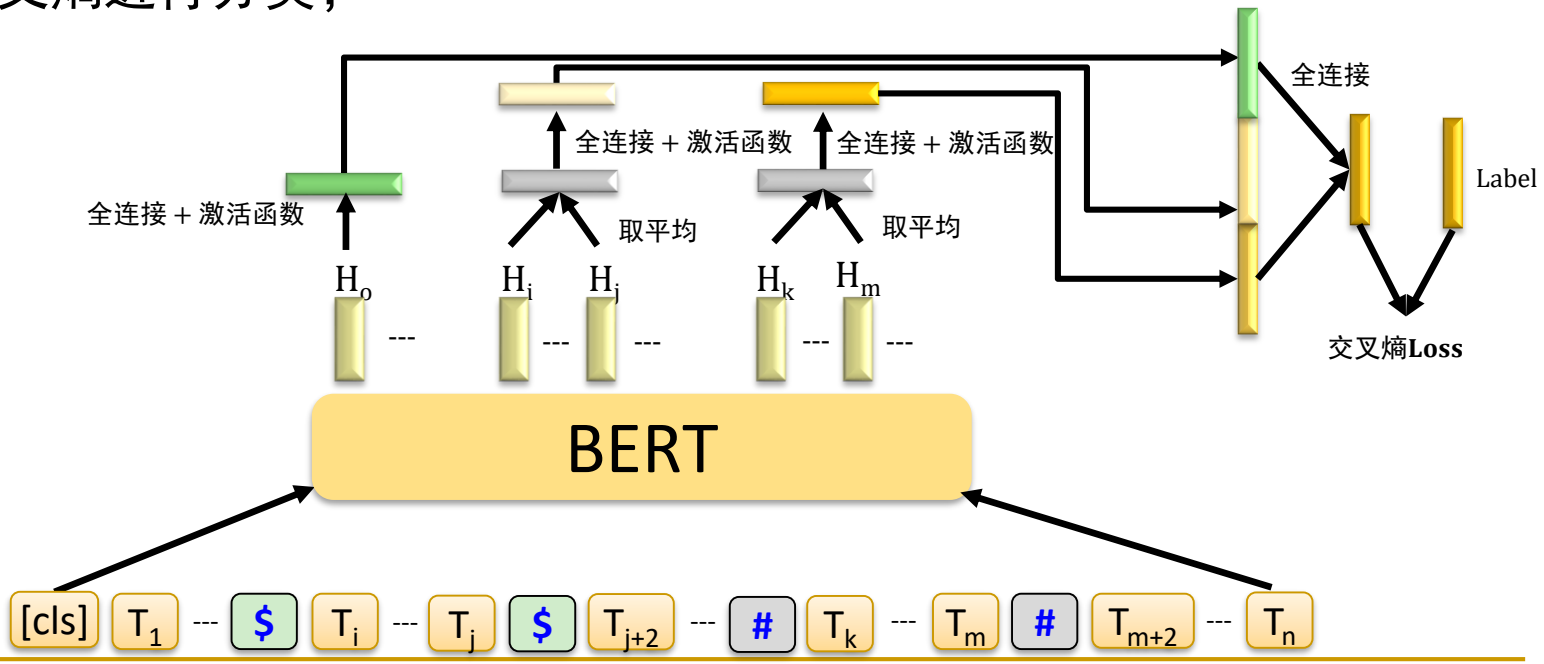
基于BERT的关系抽取模型R-BERT

- 基本框架
  - 为了能够定位两个待抽关系的目标实体并将其信息转移到BERT中，在将整个句子输入BERT前，在目标实体前后添加token，即$和#；
  - 其次通过BERT输出目标实体对应的表示；
  - 最后利用BERT输出的[CLS]隐向量和两个目标实体的隐向量进行关系分类；



基于BERT的关系抽取模型R-BERT

- 模型结构
  - [CLS]表征：该部分为单一向量，因此直接输入前馈神经网络中；
  - 实体信息：将每个实体对应的词向量进行求和平均，输入前馈神经网络中；
  - 分类：三个部分的输出向量进行拼接并输入全连接层中，最后通过交叉熵进行分类；



什么是知识推理?

- 人类视角
  - 人们从已知的事实出发，通过运用已掌握的知识，找出其中蕴含的事实或归纳出新的事实的过程
  - 按照某种策略由已知判断推出新的判断的思维过程
  - 基于特定的规则和约束，从存在的知识获得新的知识

- 计算机视角
  - 在计算机或智能系统中，模拟人类的智能推理方式，依据一定的推理控制策略，利用形式化的知识进行机器思维和求解问题的过程

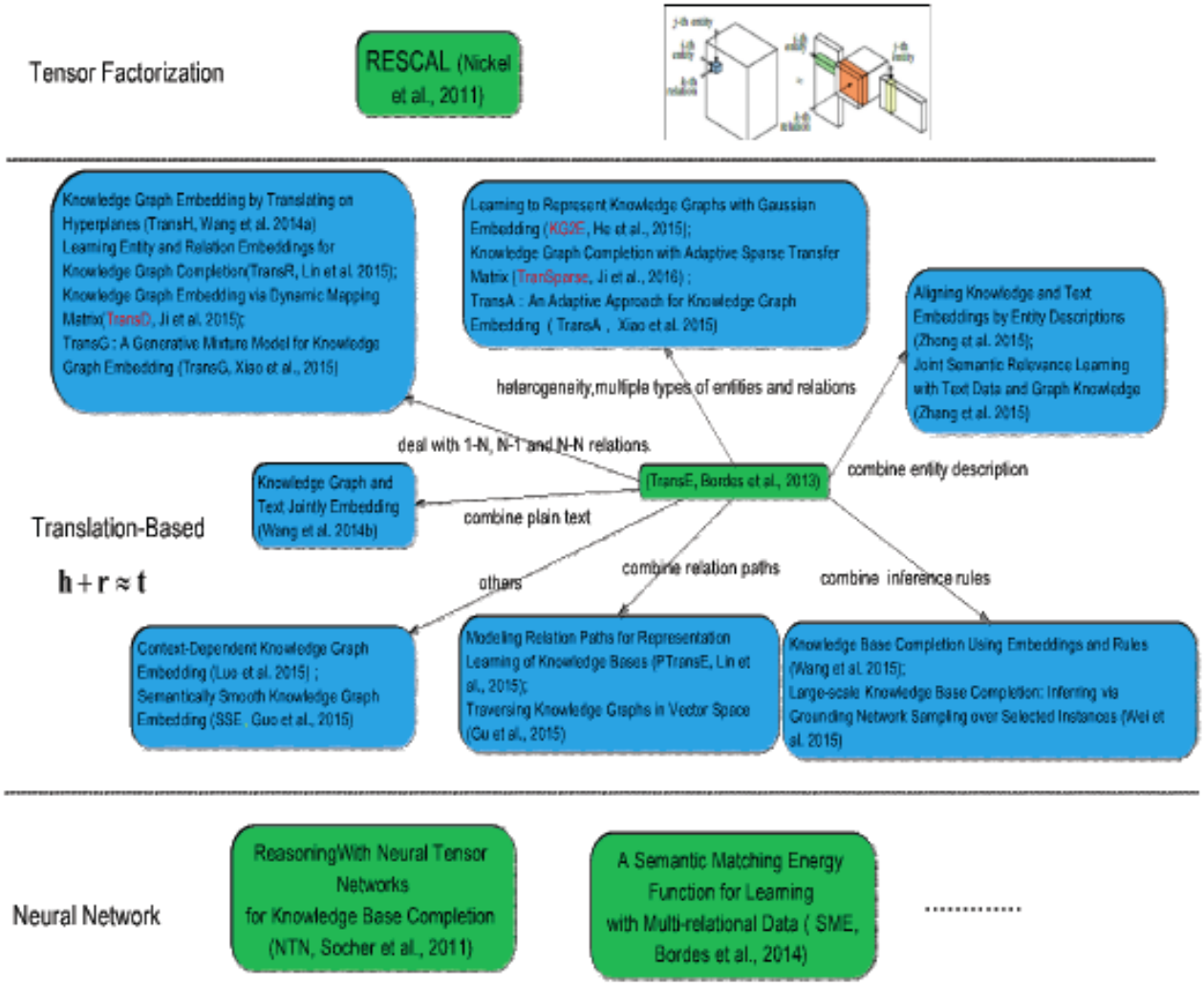
利用已知的知识推出新知识的过程

基于分布式表达的知识计算

张量分解方法

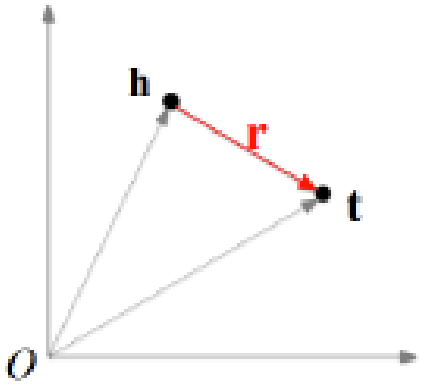
基于翻译的方法

神经网络方法



基于分布式表达的推理：TransE

- 关系事实=(head, relation, tail)，其对应的向量表示为 $(\mathbf{h}, \mathbf{r}, \mathbf{t})$
- 基本思想：把关系看作是头尾实体之间的平移(翻译)操作



$(\mathbf{h}, \mathbf{r}, \mathbf{t})$

$\mathbf{h} + \mathbf{r} = \mathbf{t}$

向量加法的三角形法则：

中国 + 首都 = 北京

法国 + 首都 = 巴黎

俄罗斯 + 首都 = 莫斯科

Bordes, et al. Translating embeddings for modeling multi-relational data. In Advances in Neural Information Processing Systems, 2013 (pp. 2787-2795).

基于分布式表达的推理：TransE

- 势能函数
  - 对于真实事实的三元组(h, r, t)，要求$\mathbf{h}+\mathbf{r}=\mathbf{t}$；而对于错误的三元组则不满足该条件

$$f(h,r,t)=f_r(h,t)=\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_2$$

$$f(\text{姚明},\text{出生于},\text{北京})>f(\text{姚明},\text{出生于},\text{上海})$$

基于分布式表达的推理：TransE

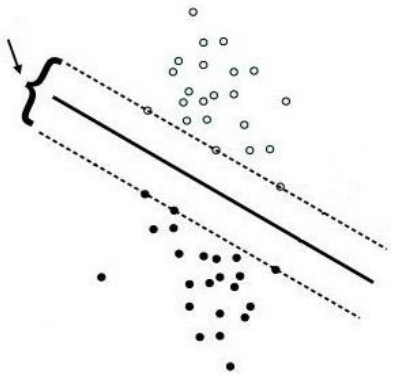
- 损失函数：

$$L = \sum_{(h,r,t) \in \triangle} \sum_{(h',r,t') \in \triangle'} \max(0, f_r(h,t) + M_{opt} - f_r(h',t'))$$

正例三元组集

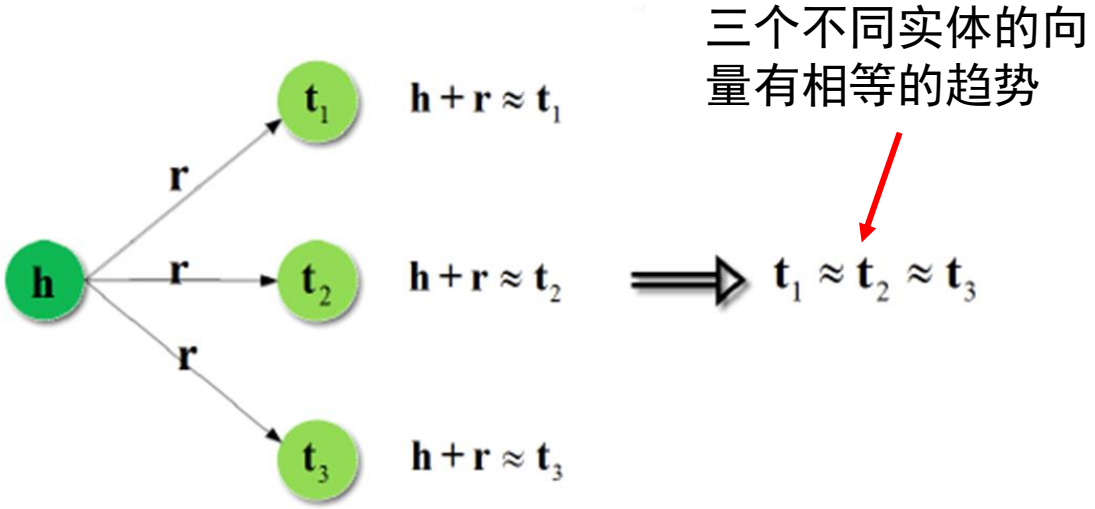
负例三元组集

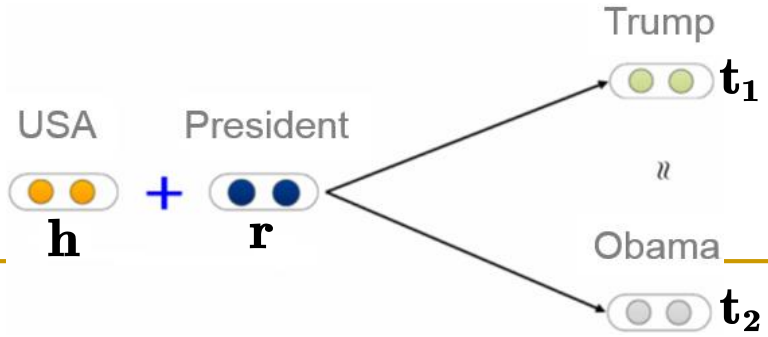
最优Margin超参



关系多样性问题

- 知识图谱中关系有1-1、1-N、N-1、N-M多种类型

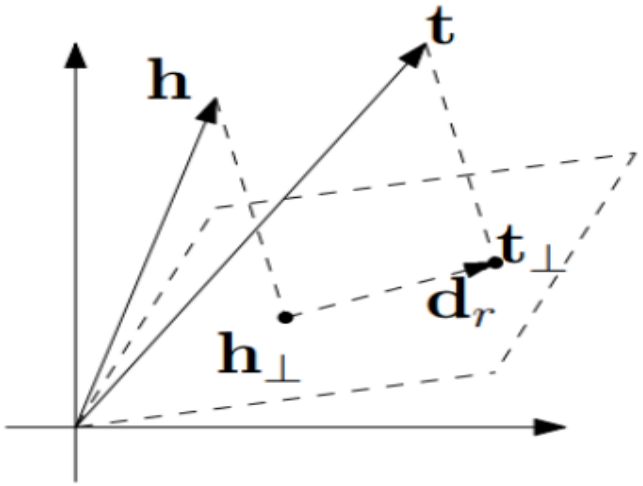




Trans系列

- TransH

$$f_r(h,t) = \|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|_{L_1/L_2}$$

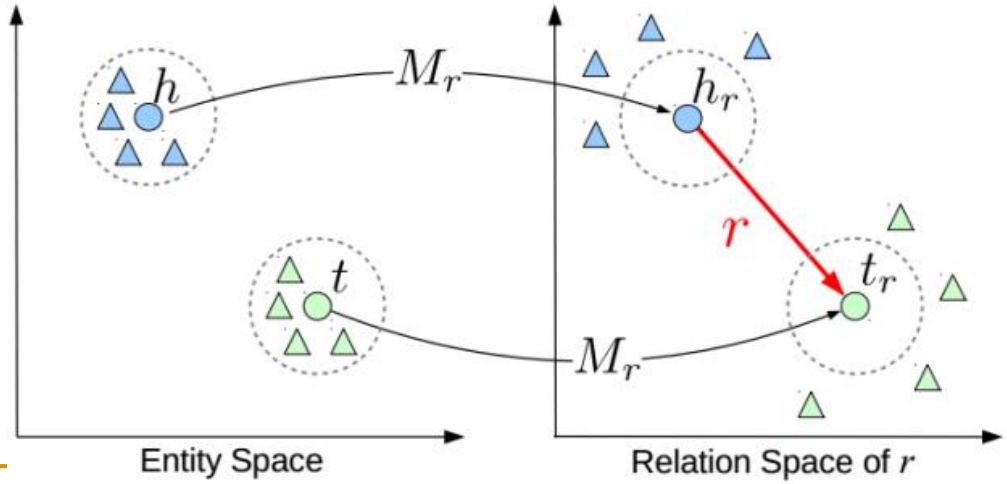


TransH

- TransR

$$f_r(h,t) = \|\mathbf{h_r} + \mathbf{r} - \mathbf{t_r}\|_{L_1/L_2}$$

- ...



TransR

谢谢！

祝大家考试取得好成绩！