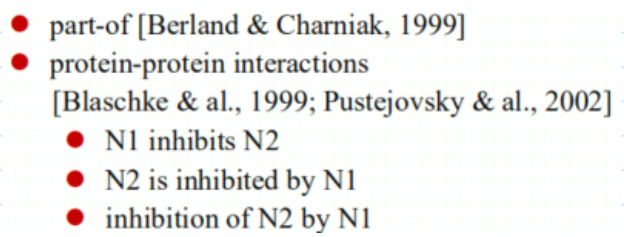
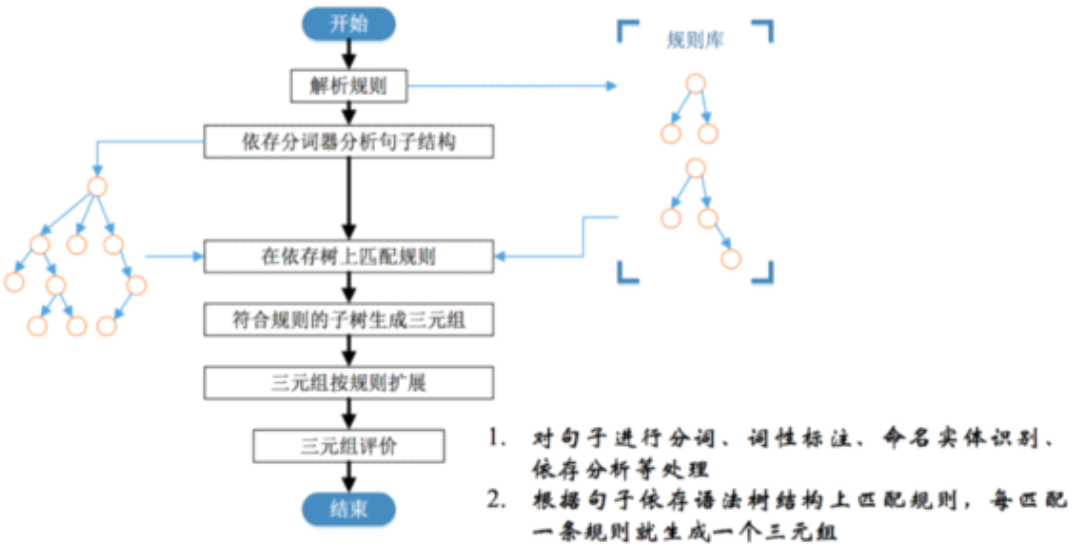
SEU 知识抽取-关系抽取b

2019年11月22日 13:22

- **基于模板的实体关系抽取**
  - **基于模板的方法**
    - 使用模式（规则）挖掘关系，基于触发词/字符等
    - 基于依存语句
  - **关系挖掘模式**
    - 支持大多数关系抽取系统的基本概念是关系模式
    - 他是一个表达式，当与文本片段相匹配时，他能够标识出相应的关系实例
    - 例如
      - 词典项
      - 通配符
      - 词性
      - 句法关系
      - 正则表达式中的灵活规则

    - **Hearst（1992）的模式清单**
      - Hearst提出了一个开创性的模式清单，用于提取分类关系is-a的实例
      - 该方法准确率高，但是召回率低
      - 仅仅包含了is-a关系
      - 之后被扩展到了其他关系上例如.
      - 
        - part-of [Berland & Charniak, 1999]
        - protein-protein interactions
        [Blaschke & al., 1999; Pustejovsky & al., 2002]
          - N1 inhibits N2
          - N2 is inhibited by N1
          - inhibition of N2 by N1
      - 这种模式设计是否可以适用于所有的关系暂时还是不清楚的

  - 基于依存语法
    通常可以以动词为起点构建规则，对节点上的词性和边上的依存关系进行限定。



董卿现身国家博物馆看展优雅端庄大方

依存分析结果

词顺序	词	词性	依存关系路径	依存关系
0	董卿	人名	1	定语
1	现身	动词	-1 Root — 一般是谓语动词	核心词
2	国家博物馆	地名	1	宾语
3	看	动词	1	顺承
4	展	动词	3	补语
5	优雅	形容词	7	定语
6	端庄	形容词	7	定语
7	大方	形容词	4	宾语

规则抽取结果

(董卿，现身，国家博物馆) ⇒ 位于(董卿，国家博物馆)

- **基于模板的实体关系抽取**
  - **优点**
    - 人工规则有高准确率（high-precision）
    - 可以为特定领域定制（tailor）
    - 在小数据集上容易实现，构建简单
  - **缺点**
    - 低召回率（low-recall）
    - 特定领域的模板需要专家构建，要考虑周全所有可能的pattern很难，很费时间精力
    - 需要为每一条关系来定义pattern
    - 难以维护
    - 可移植性差

- **有监督实体关系抽取方法**
  - **关系学习的算法**
    - **基于特征向量的方法**
      - 从上下文信息、词性、语法等中抽取一系列特征
    - **核分类**
      - 关系特征可能拥有复杂的结构
    - **序列标注方法**
      - 关系中参数的跨度是可变的
  - **基于特征向量的方法**

    定义：从上下文信息、词性、语法等中抽取一系列特征，来训练一个分类器（如朴素贝叶斯、支持向量机、最大熵等），然后完成关系抽取。

    对于一组训练数据：

    $$(x^1, y^1), (x^2, y^2), \ldots (x^n, y^n)$$

    将二元关系抽取视为分类问题：

    $$y^i = \{-1, 1\}$$

    进而学习出一个分类函数f：

    $$f(x) = \begin{cases} 1 & \text{如果}x\text{中的实体对具有某种语义关系} \\ -1 & \text{其余情况} \end{cases}$$

  - **核分类**
    - 观点：两个实体的相似度可以在高维的特征空间中计算得到而不需要枚举特征空间的各个维度
    - convolution kernels：易于对特征进行组合，例如：实体关系
    - kernelizable classifiers：SVM、Logistic Regression、KNN、Naïve Bayes
  - **Sequential labelling methods**

    一个好的关系抽取系统
    - 能够识别出句子中的实体，并且打上对应的语义类型标签。如：person、organization、location、protein、disease等
    - 对于识别出的实体，给出句子中存在的关系。如：president-of、born-in、

cause、side-effect

- HMMs、MEMMs、CRFs
- **useful for**
  - argument identification
    - e.g. born-in holds between Person and Location
  - relation extraction
    - argument order matter for some relations
- 在某些特殊情况下，关系抽取可以退化为序列标注的问题
- 例如对人物传记的百科全书文章，对于其中提到的主要实体，其他的实体都是和他相关的。因此只需要找出其他相关实体与这个主要实体之间的关系即可。此时没有必要枚举所有的实体对，因为关系是二分类的且其中的一个实体已经给出了。

- **弱监督实体关系的抽取**

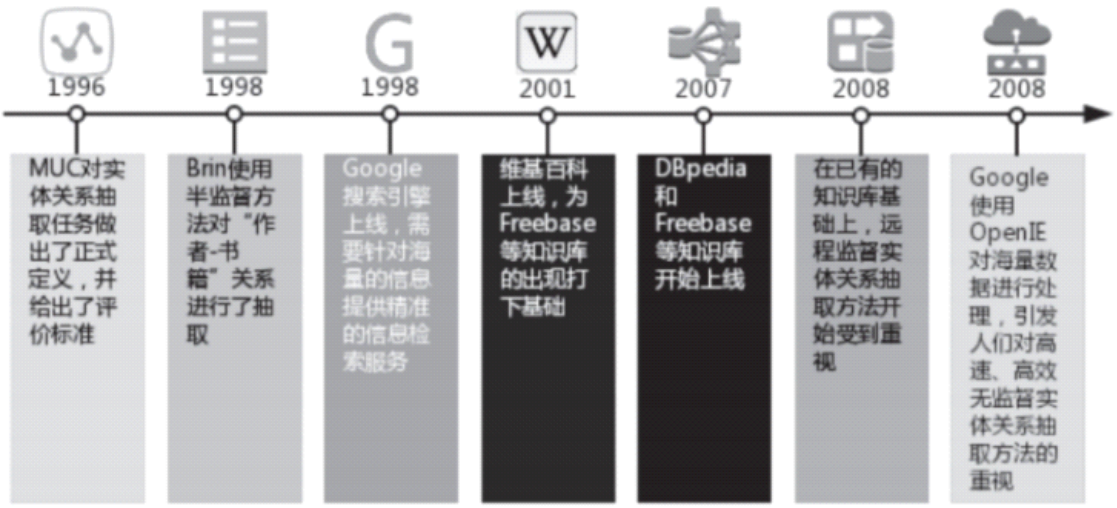


图 1　弱监督学习发展历程中的关键节点

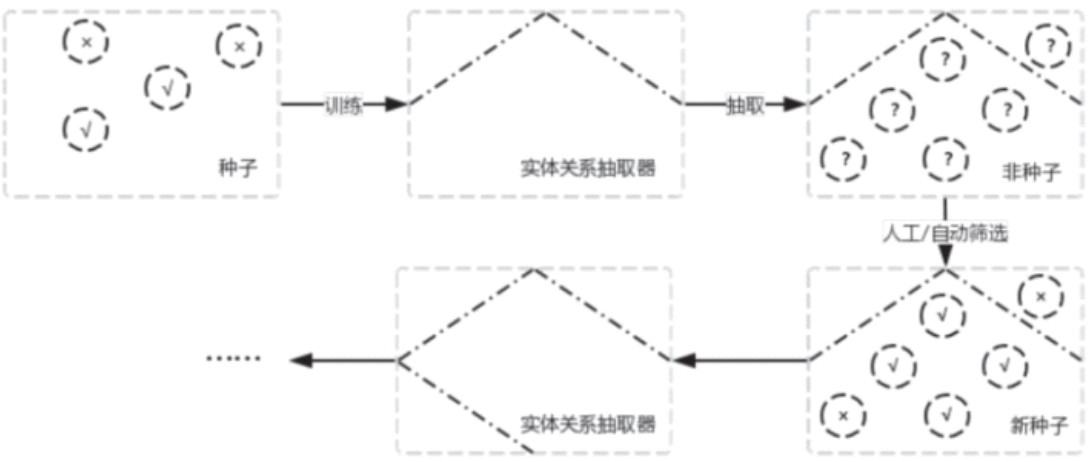


图 2　半监督学习训练过程

- 为关系抽取生成大量的标记数据是耗费时间、精力和财力的。

- **弱监督技术的出现的原因有两个：**
  - 减少构建标记数据所需要耗费的人力
  - 充分利用比较容易获得的无标记的数据

Bootstrapping方法

- 初始化
  - 给定少量的种子，
    - 该种子要具有高精度的模式P或者是要学习的关系的已知实例R
  - 例如对于is-a而言：
    - cat-animal
    - car-vehicle
    - banana-fruit
- 扩张阶段：
  - 新的模式
  - 新的实体
- 设置迭代次数
- 主要的难点
  - 语义漂移

Bootstrapping方法

- 上下文依赖
  - Bootstrapping对于上下文依赖的关系不友好
    - 在一篇报纸中： “Barcelona defeated Real Madrid.”
    - 几个月后报纸中： “Real Madrid defeated Barcelona.”
- 特别性
  - Bootstrapping对于特定关系表现很好，如birthdate
  - 无法区分细粒度的关系
  - 例如不同类型的Part-Whole：
    - Component-Integral_Object, Member-Collection, Portion-Mass, Stuff-Object, Feature-Activity and Place-Area
    - 它们可能享有相同的模式

语义漂移示例：

Seeds		Patterns		Added examples
London Paris New York	→	mayor of X lives in X ...	→	California Europe ...

策略：

- 限制迭代的次数
- 在每次迭代中选择少量的模式和实例进行添加
- 使用语义类型，例如SNOWBALL系统：
  - <Organization>'s headquarters in <Location>
  - <Location>-based <Organization>
  - <Organization>, <Location>
- 参数类型检查

策略：

对于被抽取到的模式和关系实例，在添加前会先进行打分，只有最高分的会被选中加入

- [Curran et al., 2007]提出特异性评分.

$$specificity(p) = -\log(\Pr(X \in MD(p)))$$

P是将被打分的模式， MD(p)是文本集合D中满足模式P的元组， x是一个随机变量，均匀分布在目标关系的元组域上。

- Agichtein and Gravano [2000]提出了基于准确率的评分，也是该模式的置信度：

$$Conf(p) = \frac{p.positive}{p.positive + p.negative}$$

- Pantel and Pennacchiotti [2006]根据模式和关系的可靠性定义了关系模式的置信度：

$$r_\pi(p) = \frac{\sum_{i \in I} \frac{pmi(i,p)}{\max_{i,p} pmi(i,p)} r_\iota(i)}{|I|}$$

$$r_\iota(i) = \frac{\sum_{p \in P} \frac{pmi(i,p)}{\max_{i,p} pmi(i,p)} r_\pi(p)}{|P|}$$

Label Propagation

- 2006 年，Jinxiu Chen等[42]在 ACL 会议上提出标注传播算法(Label Propagation)，这是一种基于图的弱监督学习方法。
- 基于图的学习方法都是建立在这样的假设基础上：具有相同特征的两个节点倾向于属于同一个类别.
- 而关系抽取任务的假设前提则是:如果两个关系实例相似度很高,即特征集合相似且语法结构相似,则它们将倾向于属于同一种关系类型.
- 可以看出,关系抽取任务的假设前提与基于图的学习方法的假设是吻合的.
- 因此,可以利用图来建立关系抽取模型,然后利用少部分有标签的数据辅助大量未标签的数据进行非监督的学习.

Label Propagation

方法：

- 将所有的实体对看作是图上的节点，将实体对间的距离看作边。
- 把一部分标注好的节点看作源头向其他节点传播，而权重值越高的边上传播的速度越快。
- 将相似度高的节点聚为一类， 类别信息通过传播过来的标注信息来判别。

$$W_{ij} = \exp\left(-\frac{s_{ij}^2}{\alpha^2}\right).$$

边的权重计算公式
sij代表样本节点 xi和 xj之间的相似性

$$T_{ij} = P(j \to i) = \frac{w_{ij}}{\sum_{k=1}^{n} w_{kj}}.$$

Tij 表示从节点 xj 跳到节点 xi 的概率

- **协同学习**
  - 2011 年，A.Cvitas[43]提出了一种新的弱监督方 法——协同学习(Co-learning)方法
  - 基本流程：
    - 选择 2 个不同的分类器，
    - 使用相互独立的特征在 2 个训练集上训练，并分别在未标注集上测试
    - 选取置信度高的实例扩展到另一个分类器的训练集中
    - 如此迭代若干次，当精度达到阈值时停止。

- **远程监督实体关系抽取**
  - Distant supervision for relation extraction without labeled data
    - **训练阶段**
      - 使用NET（named entity taggger）标注persons、organizations和loctions
      - 对在freebase中出现的实体对提取特征，构造训练数据
      - 训练多类别逻辑斯特回归模型
    - **测试阶段**
      - 使用NET（named entity taggger）标注persons、organizations和loctions
      - 在句子中出现的每对实体都被考虑作为一个潜在的关系实例，作为测试数据
      - 使用训练后的模型对实体对分类
    - 假设度与所有上下文中有一定比例的实例是正例，这个比例对于不

同的关系可能会不同

- **建模过程：**
  - 扩展Hoffmanm等人的图形模型以强制执行百分比约束
  - 使用感知器进行更新训练
  - 通过交叉验证使用网格搜索来找到关系的最佳百分比

- **无监督实体关系抽取**
  - **出现的原因**
    - 由于有监督和半监督机器学习方法需要事先确定关系类型，而实际上在大规模语料中，人们往往无法预知所有的实体关系类型。
    - 有些研究者基于聚类思想，利用无监督机器学习的方法，尝试解决这个问题
  - **一般过程：**
    - 实体对聚类：首先采用某种聚类方法将语义相似度高的实体对聚为一类
    - 和关系标记：在选择具有代表性的词语来标记这类关系

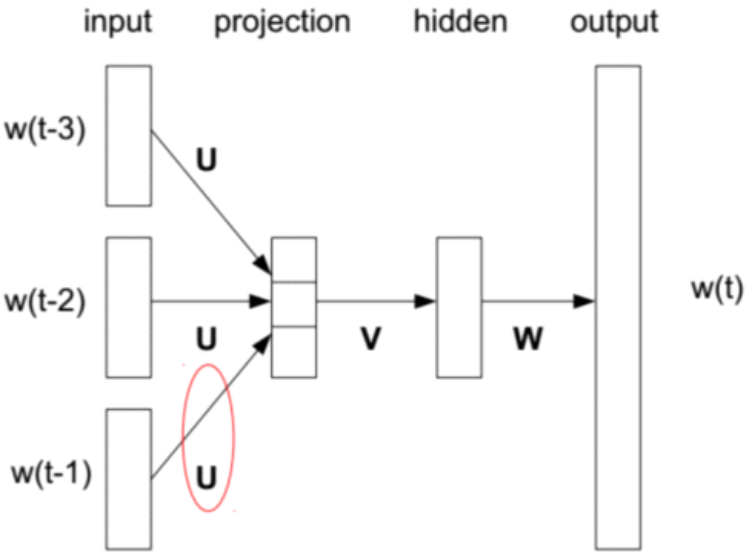
  - **一些无监督关系抽取的参考方法**
    - Discovering Relations Among Named Entities from Large Corpora 【2004】：
      - 无监督的关系抽取方法最早在2004年的ACL会议上提出，之后的方法多是在Hasegawa的基础上改进来的
    - Preemptive Information Extraction using Unrestricted Relation Discovery【2006】
      - 一种用于多聚类的无监督抽取方法
      - 首先用爬虫获取新闻文本
      - 根据文章描述进行分类
      - 再根据句子的语义结构，提取满足一系列约束的基本模式聚类实体，这些实体是根据基本模型映射来的，行程二次聚类，因此每个耳机聚类包含机油相同的实体的关系。
    - 2014 Quan提出了一种基于**模式聚类**和**句子解析**的无监督方法来处理生物医学关系抽取
      模式聚类算法基于多项式核方法，通过无标签数据识别交互词，然后将这些交互词用于实体对之间的关系抽取
    - URES
      - URES是一种主要的基于非聚类的无监督关系抽取系统
      - 他可以完全无监督从网页中抽取关系
      - 输入要求：感兴趣的关系的名字、相应的属性类型和一些关系实例的种子

- 例如：为了抽取关系“Acquisition” 你可以设置输入为 acquired acquisition

- 基于深度学习的关系抽取
  - 词嵌入
    - 映射单词到一个实值的低维空间向量
  - 如何做
    - neural networks
    - dimensionality reduction
    - explicit representation
  - 为什么关注
    - 词嵌入对许多NLP任务都非常重要 包括关系抽取

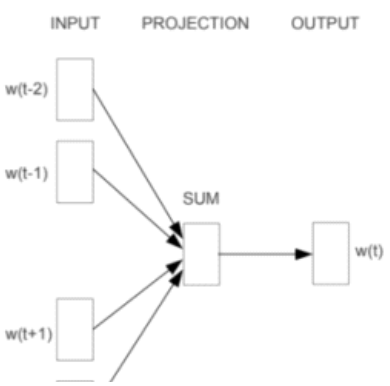
Word Embeddings from a Neural LM [Bengio &al.2003]

NNLM模型：



Efficient Estimation of Word Representations in Vector Space [Mikolov &al.2013]

本文是 word2vec 的第一篇, 提出了大名鼎鼎的 CBOW 和 Skip-gram 两大模型.



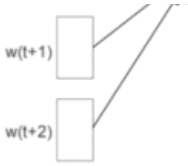
CBOW 的架构如做图所示.

与作者提到的 feedforward NNLM 的架构input->linear->hidden->output很像.

作者发现模型复杂性主要来自 non-linear hidden layer, 因此去掉了它.

由于 linear layer 被所有 words 共享, 所有 words 都会被映射到同一片空间.

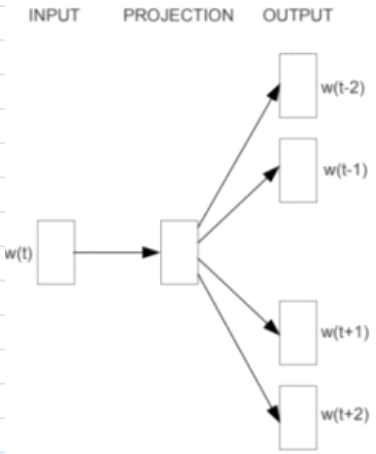
CBOW 所要做的就是利用一个单词前后的单词, 称为其上下文 context来预测它



CBOW

words 都会被映射到同一片空间.

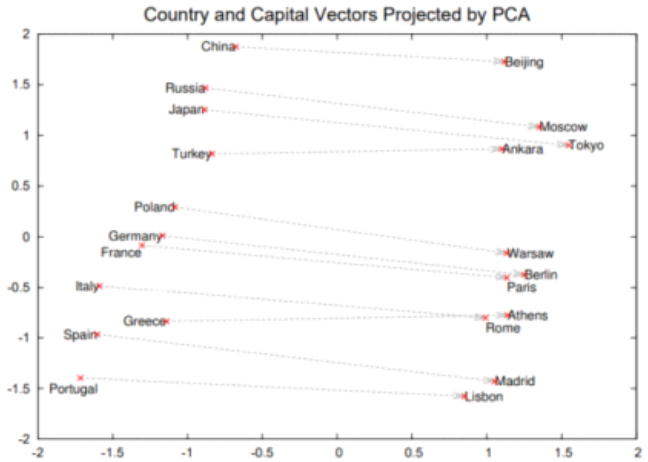
CBOW 所要做的就是利用一个单词前后的单词, 称为其上下文, context来预测它.



Skip-gram

Skip-gram 模型如下所示. 看上去与 CBOW 正好相反, 它接受一个单词作为输入, 预测在其上下文中的单词们. 文中提到, 增大 context window 能提升 word vector 的质量.

本文是 word2vec 的第二板斧, 提出了一些提升的技巧



Skip-gram: projection with PCA

Skip-gram的性质：

词嵌入具有线性结构，可以实现与矢量算术的类比

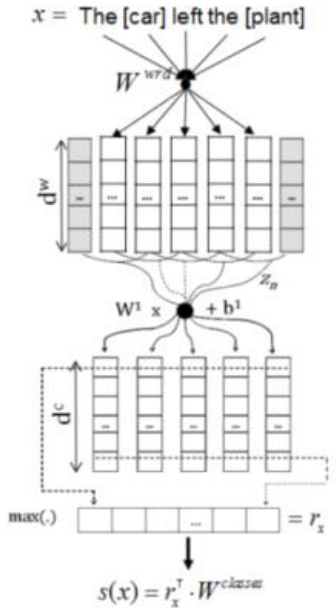
由于训练目标，输入和输出（在softmax之前）是具有线性关系的s

Skip-gram:向量计算

- 来自矢量问题的启发

Expression	*Nearest token*
Paris - France + Italy	Rome
bigger - big + cold	colder
sushi - Japan + Germany	bratwurst
Cu - copper + gold	Au
Windows - Microsoft + Google	Android
Montreal Canadiens - Montreal + Toronto	Toronto Maple Leafs

Classifying relations by ranking with convolutional neural networks. [dos Santos&al., 2015]



输入层: 利用word embedding + position embedding,同zeng 2014

卷积层: 固定尺寸的卷积核(window-size=3)

Pooling层: 直接Max Pooling得到 r_x

全连接层: 得到每个类别的score，其中 $W^{classes}$ 的每一列可以看成 label 的 embedding。因此某个label为c的score即为:

$$s_\theta(x)_c = r_x^T[W]_c$$

创新：

最大的变化是损失函数，不再使用softmax+cross-entropy的方式，而是margin based的ranking-loss

Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. [Zhou, 2016, ACL]

模型一共包括**5层结构**:

- 输入层：将句子输入到模型中
- Embedding层：将每个词映射到低维空间
- LSTM层：使用双向LSTM从Embedding层获取高级特征
- Attention层：生成一个权重向量，通过与这个权重向量相乘，使每一次迭代中的词汇级的特征合并为句子级的特征。
- 输出层：将句子级的特征向量用于关系分类，使用softmax分类器