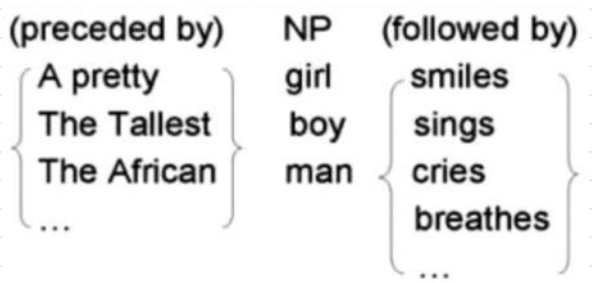
SEU 知识抽取-关系抽取

2019年11月17日 15:08

- **关系抽取简介**
  - **语义关系**
    - 是指隐藏在句法结构后面又词语的语义范畴建立起来的关系
    - 在句子中低维很重要
    - 链接文本中的实体
    - 与实体一起表达文本中的含义
    - 并不是很难识别
  - **句法关系**
    - **位置关系**
      - 位置关系是组合关系一个方面的表现
      - 也叫作横向关系
      - Word Order
        - SVO、VSO、SOV、OVS、OSV、VOS
        - 英语使用的就是SVO
    - **替代关系**
      - 指的是在某个结构的位置上彼此可以相互替换的成分之间的关系
      - the _____smiles
        - example：man boy girl
        - 可以填入词必须满足以下条件
        - 有生命的
        - 只有人类才能很自然的与smiles这个动词链接
        - 与smiles连接的一定是单数
      - 他也被称为联想关系（associative relations 索绪尔）or聚合关系
      - 为了方便理解，它也被称为纵聚合关系或纵向关系

例如，索绪尔在音位上的应用：

- 在单词pit中，/p/与/i/、/t/之间是组合关系
- 在单词pit、hit、sit中，/p/、/h/、/s/是聚合关系

    - **同现关系**
      - 同现关系值得是小句子中不统计和关系的词语允许或要求和另一集合或类别中的词语一起组成句子或句子的某一特定成分
      - 

      - 同现关系部分术语组合关系，部分属于聚合关系
    - **语义关系的双重性**
      - **逻辑方面**：谓词
        - 用于AI以支持基于知识的表示和推理
      - **图方面**：弧形链接概念

- 在NLP中可以用以表示事实性知识
- 主要是二元关系
  - 在本体论中
  - 在IE中的目标
  - ……

□ **语义网络**

- 一种用图来表示知识的结构化方式，信息呗表达为一组节点，节点通过一组代表及的有向直线彼此连接，用于表示节点间的关系
  - 顶点是映射早文本中单词的概念
  - 边代表概念间的关系

▪ **关系抽取的用处**

□ 构建知识库

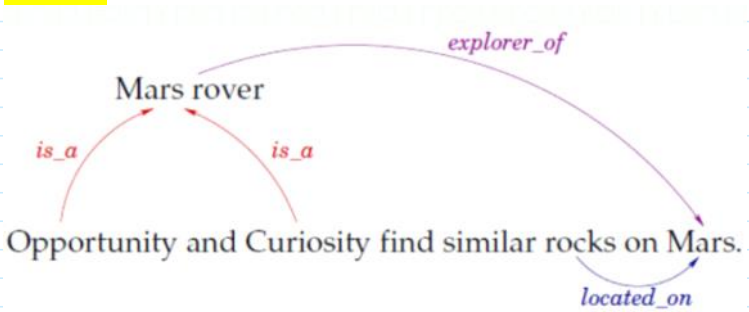
□ 文本分析

□ NLP应用

信息抽取/信息检索/自动摘要/机器翻译/问答/释义/文本蕴涵推理

叙词表构建/词义消歧/语言建模

- **语义关系**



○ **概念间的关系**

- 主要是关于世界的知识
- 可以从文本中发现

○ **名词间的关系**

- 主要是关注文本所表达的事件or形势
- 可以通过知识库信息进行发现

○ **复合名词（Noun compounds）**

定义：两个或者更多名词连在一起所构成的词

- example：silkworm/olive oil/healthcare reform/plastic water bottle

○ **复合名词的性质**

- 隐式关系编码：难以解读
- 丰富性：难以忽略
- 高产：无法被列入词典

迫切需要构建一个关系清单

- 这个清单需要有良好的覆盖率
- 关系应该是不相交的，并且各自描述一个连贯的概念
- 类分布不应过度倾斜或稀疏
- 关系背后的概念应该可以推广到其他语言现象上
- 所使用准则应该使注释过程变得尽可能简单
- 这些类别应提供有用的语义信息

摘自西格达的书籍《Semantic Relations Between Nominals》

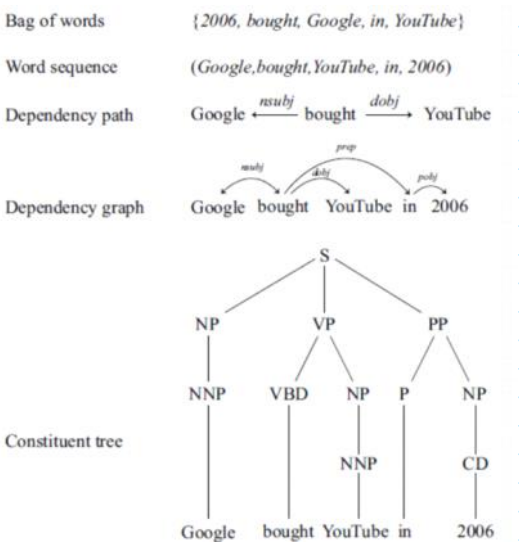
复合名词短语：使用词典释义

使用介词：存在的问题

- 介词是一词多义的：
  - school of music
  - theory of computation
  - bell of (the) church
- 没有必要的区别，例如in、on和at：
  - prayer in (the) morning
  - prayer at night
  - prayer on (a) feast day
- 一些复合词不能使用介词释义
  - woman driver
- 奇怪的释义
  - honey bee – is it “bee for honey”?

- **关系抽取中的特征**
  - **监督学习**
    - 优点：表现很好
    - 缺点：需要大量的标记数据和特征表示
  - **无监督学习**
    - 优点：可扩展，适用于开放式信息抽取
    - 缺点：表现比较差
  - **特征**
    - 目的是将数据映射为向量
    - **实体特征**
      - 捕获关系实体中的参数语义的一些表示
    - **关系特征**
      - 直接对关系进行表征，表征参数之间的相互作用，如对文本中的实体的上下文进行建模
  - **实体特征**
    - 基本实体特征包括每个候选参数的字符串值以及标记这些参数的单个单词，可能是雌性化或词干化
    - 例如：字符串值 、 单独的词、词形化或者词干化
    - **优点**
      - 在大多数情况下，这样的特征对于一个良好的关系是信息足够的
    - **缺点**
      - 这样的特征倾向于比较稀疏
  - **Basic relational features**

- 对上下文进行建模
  - 在两个参数之间的单词
  - 处于参数的特定窗口或者一侧的单词
  - 链接参数的依赖路径
  - 一个完整的依赖图
  - 最小支配子树



- **Background relational features**
  - 编码关于实体通常是如何交互的知识，而不仅仅是上下文
  - 通过事宜进行的关系表征
  - 占位符模式
  - 通过聚类寻找相似上下文
  - 用释义表征复合名词

- **关系抽取数据集**
  - 数据：MUC和ACE
  - 数据：SemEval
    - 小数量的关系
    - 标注的实体
    - 附加的实体信息（wordnet语义）
    - 句子语境+挖掘模式
    - 数据集
      - 包含七个部分，分别对应七种语义分析
      - 每个部分都有训练集和测试集==看做二分类问题
      - 搜索引擎
        - 简单的搜索模式，*in*可以搜索到Connect-Contaoner关系
        - 另一方面，结果中会有出现near misses的错误： a stitch in time
        - 监督学习的分类器可以去解决这这种问题

    - **SemEval-2007 Task4**
      - 实体间通过标签<e1> …<e1> <e2>…<e2>标注
      - 对于每一个参数，使用WordNet的语义进行手工消歧
      - Query是用来从网页中获取到响应的信息的模板
      - 评测系统必须为每一组数组打上“true” “false“的标签，也就是一个二分类的问题

名词之间的语义关系：清单

Relation	Training positive	Training size	Test positive	Test size
Cause-Effect laugh [*Cause*] wrinkles [*Effect*]	52.1%	140	51.3%	80
Instrument-Agency laser [*Instrument*] printer [*Agency*]	50.7%	140	48.7%	78
Product-Producer honey [*Product*] bee [*Producer*]	60.7%	140	66.7%	93
Origin-Entity message [*Entity*] from outer-space [*Origin*]	38.6%	140	44.4%	81
Theme-Tool news [*Theme*] conference [*Tool*]	41.4%	140	40.8%	71
Part-Whole the door [*Part*] of the car [*Whole*]	46.4%	140	36.1%	72
Content-Container the apples [*Content*] in the basket [*Container*]	46.4%	140	51.4%	74
Average	**48.0%**	**140**	**48.5%**	**78**

- SemEval-2010 Task8
  - 相比于2007中对于每一种关系提供一个单独的数据集和一个对应的二分类任务，2010仅仅提供一个单独的多类别数据集
  - 多分类任务
  - 候选的实体仍然会提供，但是评测系统需要去决策实体在关系中的槽位
  - WordNet senses和query strings将不再提供
  - 数据集中的数据量 打了很多 （超过10000条标记的句子）
  - 关系的集合变大了

关系清单

Relation	Training positive	Training size	Test positive	Test size
Cause-Effect radiation [*Cause*] cancer [*Effect*]	12.5%	1003	12.1%	328
Instrument-Agency phone [*Instrument*] operator[*Agency*]	6.3%	504	5.7%	156
Product-Producer suits [*Product*] factory [*Producer*]	9.0%	717	8.5%	231
Content-Container wine [*Content*] is in the bottle [*Container*]	6.8%	540	7.1%	192
Entity-Origin letters [*Entity*] from the city [*Origin*]	9.0%	716	9.5 %	258
Entity-Destination boy [*Entity*] went to bed [*Destination*]	10.6%	845	10.8%	292
Component-Whole kitchen [*Component*] apartment [*Whole*]	11.8%	941	11.5 %	312
Member-Collection tree [*Member*] forest [*Collection*]	8.6%	690	8.6%	233
Message-Topic lecture [*Message*] on semantics [*Topic*]	7.9%	634	9.6 %	261
Other people filled with joy	17.6%	1410	16.7%	454
Total		**8000**		**2717**

- **难点：关系清单中两组相近的关系**
  - 组1
    - Component-Whole
    - Member-Collection
    - 都是Part-Whole的特殊情况
  - 组2
    - Content-Container
    - Entity-Origin
    - Entity-Destination
    - 可以通过考虑所表达的状态是静态的还是动态的进行区分

- **关系中N-N名词复合词数据集**
  - **Nastase and Szpakowicz‘s【2003】**

    将600基础名词句子按照两种粒度进行标记，5类别的粗粒度等级如下：

将600基础名词句子按照两种粒度进行标记，5类别的粗粒度等级如下：

Relation class	# examples	Example
Causality	86	rel(nmr, wrinkle, [n, 1], laugh, [n, 2], eff)
Participant	260	rel(nmr, observation, [n, 1], radar, [n, 1], inst)
Temporality	52	rel(nmr, workout, [n, 1], regular, [a, 1], freq)
Spatiality	56	rel(nmr, material, [n, 1], cosmic, [a, 1], lfr)
Quality	146	rel(nmr, album, [n, 2], photo, [n, 1], cont)

- 缩写：
  - nmr = noun modifier relation, eff = effect, freq = frequency, inst = instrument, lfr = locationFrom, cont = container
- 格式：
  - rel(nmr,Head, [POSHead,WNSenseHead],Modifier, [POSModifier,WNSenseModifier],Relation).

- **Kim and Baldwin's【2005】**
  - 手机了来自华尔街日报的2196 noun-noun复合词
    - 1088 for training
    - 1081 for testing
  - 普通的名词 且都不是很长的复合词
  - 使用了20中关系进行标注
  - 不同于其他分离标注，kim baldwin允许多类别标注：
    - 在训练集中的94个名词复合词和测试集中的81个名词复合词被标注了多种关系

- **Tratz and Hovy 【2010】**
  - 从大型语料库和华尔街日报手机和标注了17509个名词复合词
  - 比较大的关系如下
    - ***perform/engage in* (13.24%)**
    - ***create/provide/sell* (8.94%)**
    - ***topic of communication/imagery/info* (8.37%)**
    - ***location/geographic scope of* (4.99%)**
    - ***organize/supervise/authority* (4.82%)**

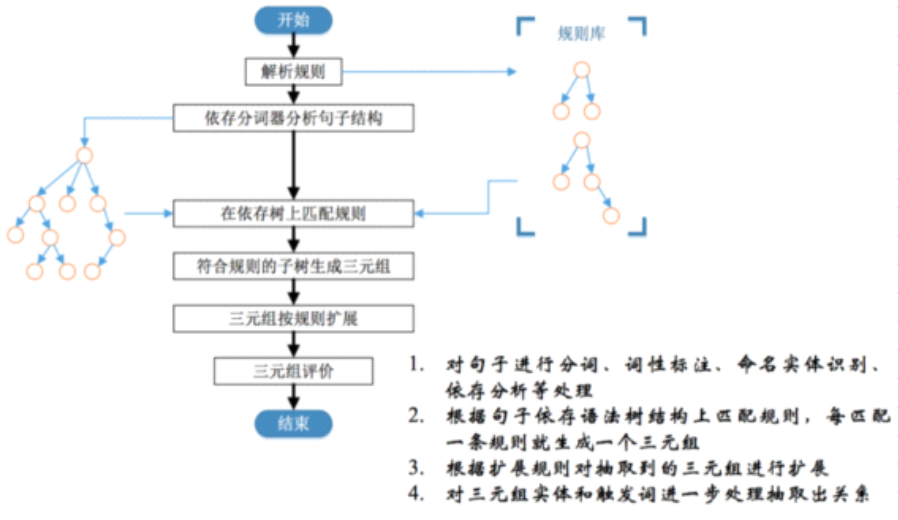
- **基于模板的实体关系抽取**
  - **基于模板的方法**
    - 使用模式（规则）挖掘关系，基于触发词/字符等
    - 基于依存语句
  - **关系挖掘模式**
    - 支持大多数关系抽取系统的基本概念是关系模式
    - 他是一个表达式，当与文本片段相匹配时，他能够标识出相应的关系实例
    - 例如
      - 词典项
      - 通配符
      - 词性
      - 句法关系
      - 正则表达式中的灵活规则

    - **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。



- **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**

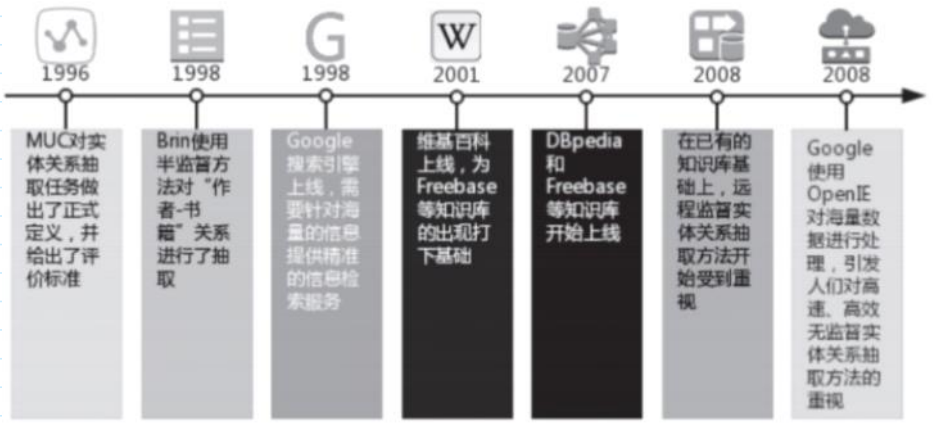


图 1　弱监督学习发展历程中的关键节点

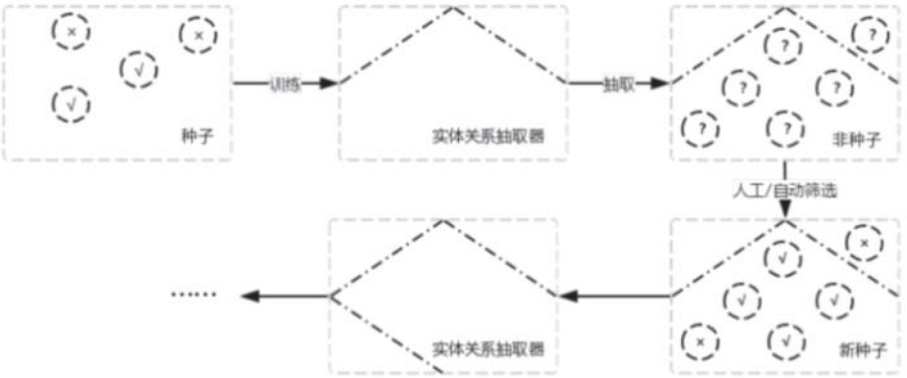


图 2　半监督学习训练过程

- 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

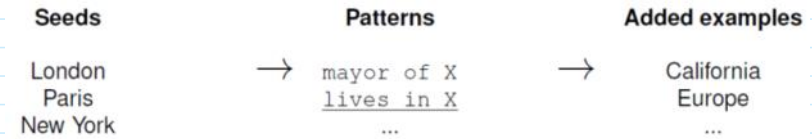
Bootstrapping方法

- 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.”

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：



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。