关系抽取

1.关系抽取的定义：从海量的非结构外的文本中抽取出有用的信息，并结构化成下游工作可用的格式，这是信息抽取的存在意义。

2. 关系抽取分两步：一步是判断一个实体对是否有关系，而另一步则是判断一个有关系的实体对之间的关系属于哪种。当然这两步可变成一步，即把无关系当作关系的一种（特殊的），来进行多类别分类。

关系抽取通常会基于以下几种方式去做：

* 基于规则
* 监督学习
* 半监督 & 无监督学习
  + Bootstrap
  + Distant Supervision
  + 无监督学习

基于规则的方法，表示 “is - a”

需要事先定义对应的规则模板，比如：

1. " … apple is a fruit … "
2. " … fruit such as apple … "
3. " … fruit including apple , banala … "

对应的优缺点比较（规则模板类方法的通病）

* 方法优点
  + 比较准确
  + 对于垂直场景，比较适合（具有针对性）
* 方法缺点
  + 信息缺乏覆盖率(low recall rate)
  + 人力成本较高
  + 很难设计（规则冲突、重叠）

监督学习的方法进行关系抽取

定义好关系的类型

定义好实体的类型

准备训练数据

实体标注

实体之间的关系标注

举例说明：**American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.**

首先我们需要运用命名实体识别(NER)工具来分析这句话，可以找到:

ORG : American Airlines

PER：Tim Wagner

然后我们需要提取特征，将(特征(American Airlines),特征(Tim Wagner))放入分类算法。  
特征提取部分可以有以下方式：

**Bag-of-words**  
例如我们可以用unigram提取出(American Airlines,Tim Wagner)，或者使用Bi-gram提取出(American Airlines a,Tim Wagner said)

**pos-tagging**  
提取出American Airlines和Tim Wagner的词性

**Bag-of-words feature**  
a ,unit, of, AMR, immediately, matched, the, move, spokesman

**between words pos-tagging**  
抽取(a ,unit, of, AMR, immediately, matched, the, move, spokesman)的词性

**head feature**  
Airlines,Wagner

**实体类别**  
(ORG,PER)

**句法分析**

**包括分析实体名词之间的间距、实体在文中的位置以及句子在文中的位置等问题。**

**Bootstrap方法**

**Bootstrap的大致思路如下：**

* **目标关系：burial place**
* **Seed tuple: [Mark Twain, Elmira]**
* **寻找包含Mark Twain, Elmira的句子**
  + **" Mark Twain is burind in Elmira, NY."  
    -> X is burind in Y**
  + **" The grave of Mark Twain is in Elmira"  
    -> TH grave of X is in Y**
  + **" Elmira is Mark Twain’s final resting place"  
    -> Y is X’s final resting place**
* **利用这些pattern去搜索更多的Tuple**

所以在传统的Bootstrap的框架下，我们引申出了snowball。

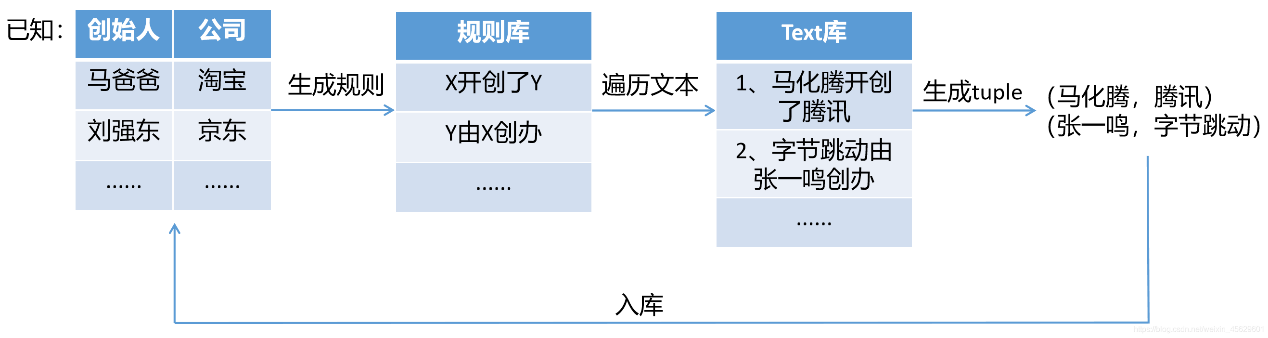
* step1 : 生成模板(Pattern)  
  比如我们有一个 [LOC] based [ORG]模板，如果文本中出现了 xxx base xxx ，此时在Bootstrap中是匹配不到的。snowball在此时做了优化，将模板部分进行向量化，然后将匹配的文本的实体间的内容，做相似度的对比。具体细节就不在这里多展开了，感兴趣的同学可以[阅读原论文](http://www.mathcs.emory.edu/~eugene/papers/dl00.pdf).
* step2: 生成Tuples  
  遍历文本数据，将相似度超过设定阈值的tuple保存下来
* step3:评估Patterns和Tuples  
  使用已有的Tuples去评估Patterns的准确率  
  使用已有的Patterns评估Tuples的准确率  
  Confidence(Tuple) = 1 - (1-Conf(pi))\* … \*(1-Conf(pk)) # 多个靠谱的Patterns的乘积

一般来讲，可以根据数据的特点将其分为结构化数据如(数据库表等)、半结构化数据如（infoxBox等）和非结构化数据（网络语料）。

在下面是详细的推导过程

**★★★半监督学习(bootstrap/Snowball)**

**bootstrap**



根据已知的seed库建立规则库，然后根据规则遍历文本，将符合规则的实体生成tuple放入seed库。最大的缺点就是error accumulate,错误的规则生成会发生传递并且不会消失，导致准确率越来越低。

还是类似于基于规则模板的方法

此外还有 使用流水线的方法（Pipelined Method）进行抽取：输入一个句子，首先进行命名实体识别，然后对识别出来的实体进行两两组合，再进行关系分类，最后把存在实体关系的三元组作为输入。流水线的方法存在的缺点有：1）错误传播，实体识别模块的错误会影响到下面的关系分类性能；2）忽视了两个子任务之间存在的关系，例如图中的例子，如果存在Country-President关系，那么我们可以知道前一个实体必然属于Location类型，后一个实体属于Person类型，流水线的方法没法利用这样的信息。3）产生了没必要的冗余信息，由于对识别出来的实体进行两两配对，然后再进行关系分类，那些没有关系的实体对就会带来多余信息，提升错误率。

6、大类：基于神经网络的

**LSTM+CRF**

最经典的 [LSTM+CRF](https://link.zhihu.com/?target=http%3A//hiheng%2520huang%252C%2520wei%2520xu%252C%2520kai%2520yu.%2520bidirectional%2520lstm-crf%2520models%2520for%2520sequence%2520tagging.%2520corr.%25202015/)，端到端的判别式模型，LSTM 利用过去的输入特征，CRF 利用句子级的标注信息，可以有效地使用过去和未来的标注来预测当前的标注。

基于注意力机制的LSTM关系抽取

https://www.ixueshu.com/document/6a3261ae85e489b8aee0181925515a37318947a18e7f9386.html

[Socher et al. 2012]提出使用**递归神经网络**来解决关系抽取问题。  
1. 该方法首先对句子进行句法解析；  
2. 然后为句法树上的每个节点学习向量表示；  
3. 通过递归神经网络，可以从句法树最顶端的词向量开始，按照句子的句法结构迭代合并，最终得到该句子的向量表示，并用于关系分类。  
该方法能够有效地考虑句子的句法结构信息，但同时该方法无法很好的考虑两个实体在句子中的位置和句义信息。

[Zeng et al. 2014]提出采用**卷积神经网络**进行关系抽取。采用**词汇向量**和**词的位置向量**作为卷积神经网络的输入，通过卷积层、池化层和非线性层得到句子表示。通过考虑实体的位置向量和其他相关的词汇特征，句子中的实体信息能够较好地考虑到关系抽取中。

后来，[Santos et al. 2015]还提出了一种**新的卷积神经网络**进行关系抽取，其中采用了**新的损失函数**，能够有效地提高不同关系类别之间的区分性。

[Miwa et al. 2016]提出了一种基于端到端神经网络的关系抽取模型。该模型使用双向LSTM和树形LSTM同时对实体和句子进行建模。  
目前基于卷积神经网络的方法在关系抽取的标准数据集SemEval-2010 Task 8 上取得了最好的效果。

上面介绍的神经网络模型在人工标注的数据集上取得了巨大成功。然而，与之前基于特征的关系抽取系统类似，神经网络关系抽取模型也面临着人工标注数据较少的问题。  
[Zeng et al. 2015]尝试**将基于卷积神经网络的关系抽取模型扩展到远程监督数据上**。其中假设每个实体对的所有句子中至少存在一个句子反映该实体对的关系，提出了一种新的学习框架；以实体对为单位，对于每个实体对只考虑最能反映其关系的那个句子。该方法在一定程度上解决了神经网络关系抽取模型在远程监督数据上的应用，在 NYT10 数据集上取得了远远高于基于特征的关系抽取模型的预测效果。但是，该方法仍然存在一定的缺陷：该模型对于每个实体对只能选用一个句子进行学习和预测，损失了来自其他大量的有效句子的信息。

[Lin et al. 2016]提出了一种基于句子级别注意力机制的神经网络模型来解决这个问题，该方法能够根据特定关系为实体对的每个句子分配权重，通过不断学习能够使有效句子获得较高权重，而有噪音的句子获得较小的权重。与之间的模型相比，该方法取得较大提升。实现代码：<http://github.com/thunlp/NRE>

认知：

实体抽取的本质是将一句话中的实体抽取出来，并根据算法、规则或者模型，得到实体键的关系，关键要点有：实体抽取、关系选择

补充：无监督关系抽取方法主要基于分布假设理论，分布假设的核心思想是：如果两个词的用法相似即出现在相同上下文中

关系抽取的最初目的是为了方便与扩充知识图谱

<https://blog.csdn.net/qq_41375318/article/details/103221279> 感觉并没有理论证明

<https://blog.csdn.net/qq_41375318/article/details/103079201> 2019年论文，方法较新

以上是两种比较好的方法

2020比较新的关系抽取论文学习：

以(**subject, relation, object**)的形式表示三元组

三元组重叠的问题：

Normal代表没有重叠的部分

EPO关系两端的实体是一致的

SEO：关系两端只有单个实体共享

**以往工作的不足以及重叠三元组出现的挑战**：

体对的组合之中，大多数实体对是没有关系链接的，这便存在很多的负例，也就造成了关系分类的不平衡。

三元组的问题更是一个难点，因为其存在共享的实体，甚至两个实体存在多种关系，这便增加了难度，没有足够的训练数据，是难以学习或者根本无法学习这种关系的。

所以，提出了新框架CASREL

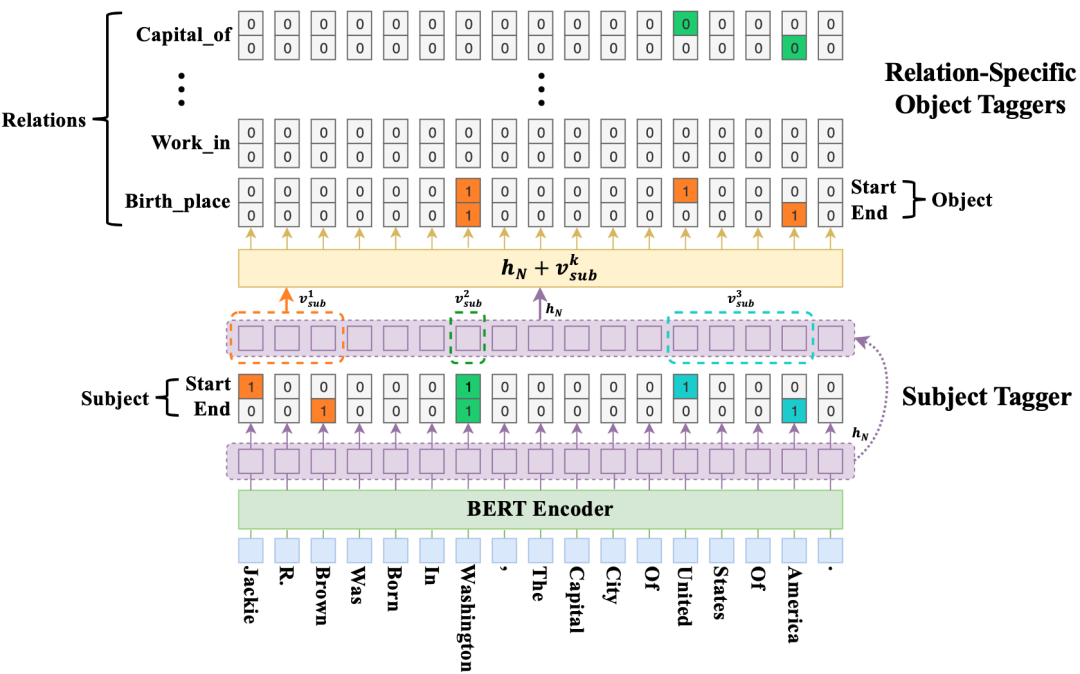
CASREL框架抽取三元组(subject, relation, object)主要包含两个步骤，三个部分。

**两个步骤：**

1. 第一步要识别出句子中的 subject 。
2. 第二部要根据识别出的 subject, 识别出所有有可能的 relation 以及对应的 object。

**三个部分：**

1. BERT-based encoder module: 可以替换为不同的编码框架，主要对句子中的词进行编码，论文最终以BERT为主，效果很强。
2. subject tagging module：目的是识别出句子中的 subject。
3. relation-specific object tagging module：根据 subject，寻找可能的 relation 和 object。
4. **其中 a 是 Encoder, b 和 c 称为 Cascade Decoder。**



整体模型如上图所示

下面是细节部分的整理

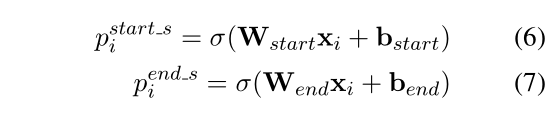
**首先**需要对句子进行Word2Vector转化，使用的技术可以是Bert的任意一层，或者是LSTM替换Bert

Bert是谷歌提供的12层的可用于NLP的深度学习模型

**接下来**

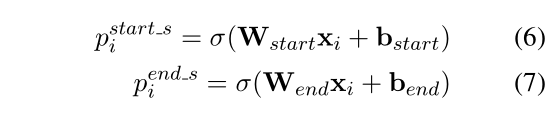
对 BERT Encoder 获取到的词的隐层表示解码，构建两个二分类分类器预测 subject 的 start 和 end 索引位置，对每一个词计算其作为 start 和 end 的一个概率，并根据某个阈值，大于则标记为1，否则标记为0。

采用的公示如下所示：



Jackie 被标记为 start，R 既不是 start 也不是 end, Brown 被标记为 end，其他的类似。在这里采用了最近匹配的原则，即与 jackie 最近的一个 end 词是 Brown, 所以 **Jackie R. Brown 被识别为一个subject**。

**Subject Tagger**这部分的主要作用是对 BERT Encoder 获取到的词的隐层表示解码，构建两个二分类分类器预测 subject 的 start 和 end 索引位置，对每一个词计算其作为 start 和 end 的一个概率，并根据某个阈值，大于则标记为1，否则标记为0。公式如下：（这部分没看懂）



如最上面框架图中所示，Jackie 被标记为 start，R 既不是 start 也不是 end, Brown 被标记为 end，其他的类似。在这里采用了最近匹配的原则，即与 jackie 最近的一个 end 词是 Brown, 所以 **Jackie R. Brown 被识别为一个subject**。文中并未考虑前面位置的情况。

没有合理的数学证明

对于feather的理解：实际上就是匹配规则模板，X founded Y || X，the founder of Y

其他可行的方法实现

增加全监督关系抽取PCNN的复现代码:

https://github.com/ShomyLiu/pytorch-pcnn

一篇关系抽取的先进论文：

Relation Classification via Convolutional Deep Neural Network

下面的链接比较详细地介绍了NLP的经典算法历程

https://blog.csdn.net/qq\_34357872/article/details/102835846?utm\_medium=distribute.pc\_relevant.none-task-blog-OPENSEARCH-2.control&depth\_1-utm\_source=distribute.pc\_relevant.none-task-blog-OPENSEARCH-2.control