1引言

传统意义上讲，关系抽取是实体识别基础上的一个任务，其核心是抽取一个句子中包含实体对之间的关系。第一步就是训练一个关系抽取器，换句话讲就是训练一个关系分类器。因为模型不可能自己给关系起名字，所以我们需要人工的标注好语料，基于语料库我们一共有多少种关系。当模型训练好了之后，给它一个包含两个实体的句子，我们通过特征提取生成一个句子向量，通过根据语料数据训练出的关系抽取器来判定该句子应当分到哪一类中去，从而完成句子中两实体的关系分类。

传统关系抽取有以下三种方法:

1. Supervised relation extraction:对句子中的关系进行人工标注，然后抽取词法、句法、语义特征，训练有监督的分类器，对测试集中的实体关系进行预测；

2. Unsupervised information extraction:从大量文本中，对实体进行标注，通过聚类等方法抽取实体之间的关系字符串。这种方法可以抽取非常大量的关系对，但是抽取结果很难映射到知识库中。

3. bootstrap learning:利用少量的种子实例或模板，进行bootstrap的学习。将已有的实例应用在大规模语料中，抽取出新的模板；然后用新的模板抽取出更多的实例，如此迭代。但抽取结果准确率较低，还可能发生语义转移。

2 远程监督

Distant Supervision，是目前关系抽取中比较常见的一类做法。该方法由 M Mintz 大佬于ACL2009上首次提出，它既不是单纯的传统意义上的监督语料，当然也不是无监督。它是一种用KB去对齐朴素文本的标注方法（Distant supervision for relation extraction without labeled data ）。

KB中已经有关系名和实体对的三元组，只需要把这三元组付给朴素文本中相应的句子就可以了，那按照什么原则付？（数据库监督）M Mintz 提出了一种十分大胆的假设:

The intuition of distant supervision is that any sentence that contains a pair of entities that participate in a known Freebase relation is likely to express that relation in some way.

即，如果我们训练语料中的句子所包含的实体对在数据库中有关系的体现，那么我们认为语料库中所有包含相同实体对的句子都表达此关系。那么，自然而然的我们提取特征的时候就会出现很多噪声。比如两个实体之间还有别的关系或者是没有关系，这样的训练数据会对我们的关系抽取器产生影响。

2.1 training

1.使用NET（named entity tagger）标注。

2.对在freebase中出现的实体对提取特征（从所有出现该实体对 的句子中），构造训练数据。

3. Multiclass logistic regression classifie

 论文中采用的NET标注工具为斯坦福的NRT标注器，再生成训练集的过程中，我们首先对大量文本句子进行命名实体标注，如果一个句子中含有两个实体，且这两个实体在Freebase（ KB）中是一个关系对，那么从句子中提取特征，将关系作为类别，直到该类别中不再有新的句子加入，我们从该类别中的所有句子提取特征向量并且合并成一个更大的特征向量，从而训练出Multiclass logistic regression classifie。

 2.2 testing

1. 使用NET（named entity tagger）标注。

2. 在句子中出现的每对实体都被考虑做为一个潜在的关系实例， 作为测试数据。

3. 使用训练好的模型对实体对进行分类。

 在测试阶段，先对句子中的命名实体进行标注，抽取其中的命名实体对和特征。如果多个句子的命名实体对一样，则将它们的特征合并在同一个特征向量中。然后利用逻辑回归分类器，对关系名称进行识别。这种方法的好处是可以综合文本中的多处，对一个实体对进行关系判断。

例如：<Steven Spielberg, Saving Private Ryan >---film-director

我们单看以下的第一个句子和第二个句子，都不能判断出Steven Spielberg和Saving Private Ryan之间存在film-director关系，但是把两个句子结合起来我们就能做到了。

 [Steven Spielberg]’s film [Saving Private Ryan] is loosely based on the brothers’ story.

 Allison co-produced the Academy Awardwinning [Saving Private Ryan], directed by [Steven Spielberg]...

3 特征提取

 Lexical features

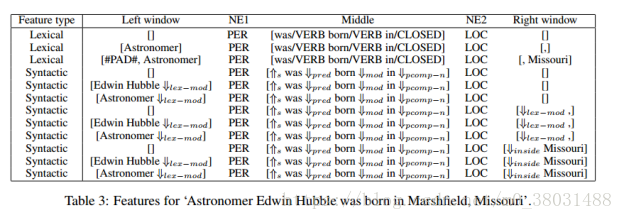
1. 实体对之间的词串;

2. 这些词的词性标记;

3. 哪个实体出现在前;

4. 实体1左边k个词和它们的POS标记;

5. 实体2右边k个词和它们的POS标记。



Syntactic features

1. 两实体之间的依存路径

2. 对每个实体，增加一个window node做为特征。window node是指与其中一个实体连接，但并不在依存路径中的节点。

论文中是结合词法特征提取、句法特征提取和命名实体特征提取。通常我们现在使用深度学习的方法来进行特征提取，所取得的效果也是有目共睹的。例如（Neural Relation Extraction with Selective Attention over Instances ），使用PCNN对句子分段进行特征提取然后句子建模，将训练句子表示成一个sentence vector，然后在进行关系分类。

4 结论

 Distant Supervision的方法虽然从一定程度上减少了模型对人工标注数据的依赖，但该类方法也存在明显的缺点：

1. 假设过于肯定，难免引入大量的噪声数据（“Steven Jobs”，“Apple ”）

2. 数据构造过程依赖于 NER 等 NLP 工具，中间过程出错会造成错误传播问题。

近些年，对于大量噪声数据的过滤的研究也一直在进行中，例如使用Multi-instance从训练集中抽取取置信度高的训练样例训练模型，利用Attention模型对数据进行全方位的权重计算，从而得到全面而不失“选择”的训练数据。对于错误传播放大的问题近期更有“joint learning”方法的提出，将命名实体识别和关系抽取两部并为一步走。原本流水线（pipeline ）式的方法也是人为的将一个句子的信息抽取分成两部走，现在这种更加“原生态”的关系抽取方式说不定能够大放异彩。