来源：Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks

# 背景知识：Distant Supervised Relation Extraction

该方法由 M Mintz 于 ACL2009 上首次提出，与传统预先定义关系类别不同，Distant Supervision通过将知识库与非结构化文本对齐来自动构建大量训练数据，减少模型对人工标注数据的依赖，增强模型跨领域适应能力。

Distant Supervision的提出主要基于以下假设：

两个实体如果在知识库中存在某种关系，则包含该两个实体的非结构化句子均能表示出这种关系。例如，"Steve Jobs", "Apple"在 Freebase 中存在 founder 的关系，则包含这两个实体的非结构文本“Steve Jobs was the co-founder and CEO of Apple and formerly Pixar.”可以作为一个训练正例来训练模型。这类数据构造方法的具体实现步骤是：

1. 从知识库中抽取存在关系的实体对
2. 从非结构化文本中抽取含有实体对的句子作为训练样例

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。如 "Steven Jobs passed away the daybefore Apple unveiled iPhone 4s in late 2011."这句话中并没有表示出 Steven Jobs 与 Apple 之间存在 founder 的关系。
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。针对这些问题，目前主要有四类方法：  
   （1）在构造数据集过程中引入先验知识作为限制；  
   （2）利用指称与指称间关系用图模型对数据样例打分，滤除置信度较低的句子；  
   （3）利用多示例学习方法对测试包打标签；  
   （4）采用 attention 机制对不同置信度的句子赋予不同的权值。

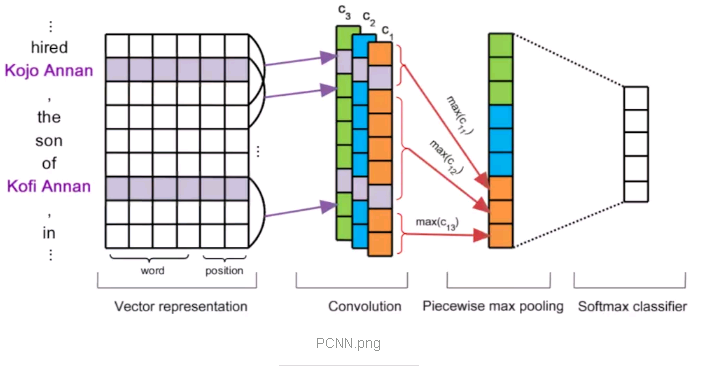
# 论文正文：

这篇论文主要针对数据标注错误问题和传统统计模型特征抽取过程中出现的错误提出解决方法。

* 针对数据标注错误问题，作者提出采用多示例学习的方式从训练集中抽取取置信度高的训练样例训练模型。
* 针对第二个问题，作者提出 piece-wise 的卷积神经网络（PCNN）。

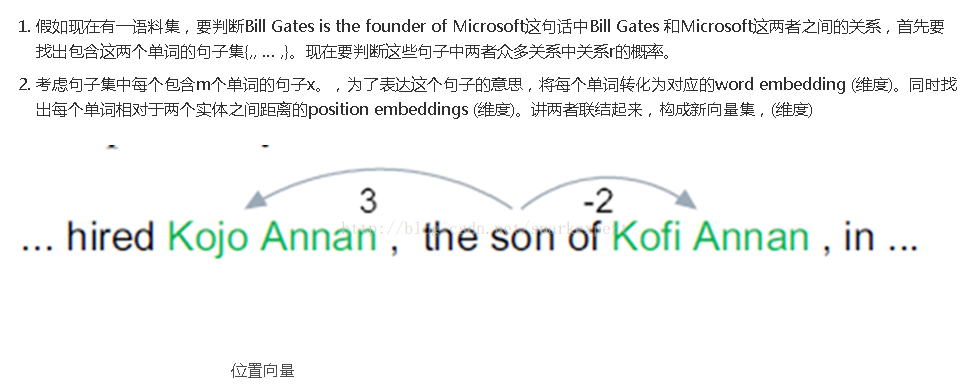
传统的关系抽取方法，包含:bootstrap、无监督发现或者有监督分类集中模式。其中有监督的方法是最常使用的，也是性能表现最好的一类，有监督主要是将关系转换成类别标签去考虑。但是，有监督的方法需要大量的数据，如果基于人工去标注每一句话当中的距离，那么明显是不现实的，也因此，这个方法也一直受制于训练数据量不足的问题。为了解决这个问题，才引入了Distant Supervised Relation Extraction，引入外部KB来实现自动化的语料标注。自动标注解决了语料的数量问题，但是其标注质量一般，存在大量的wrong label，于是又引入了multi-instance learning来解决这个问题。与此同时，如何提取特征也是一个很重要的工作，主流的方法有两种，基于特征的，无论基于什么方法，都难以找到高质量的特征。  
于是，本文准备使用PCNN进行特征抽取和分类。

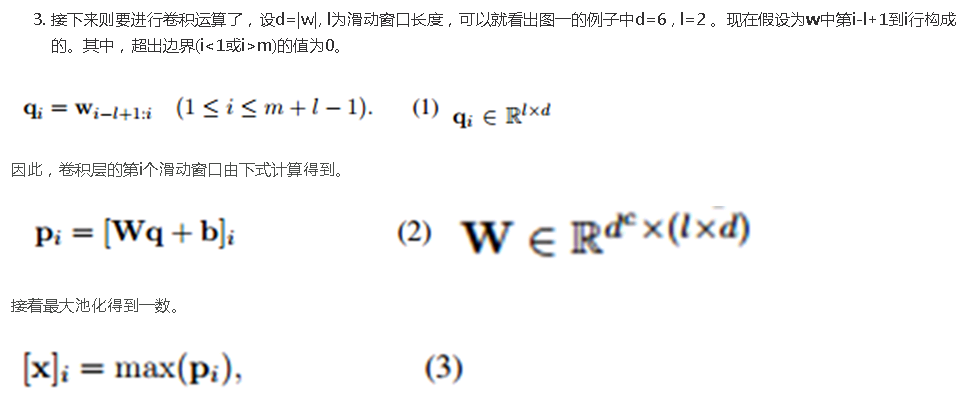
# PCNN

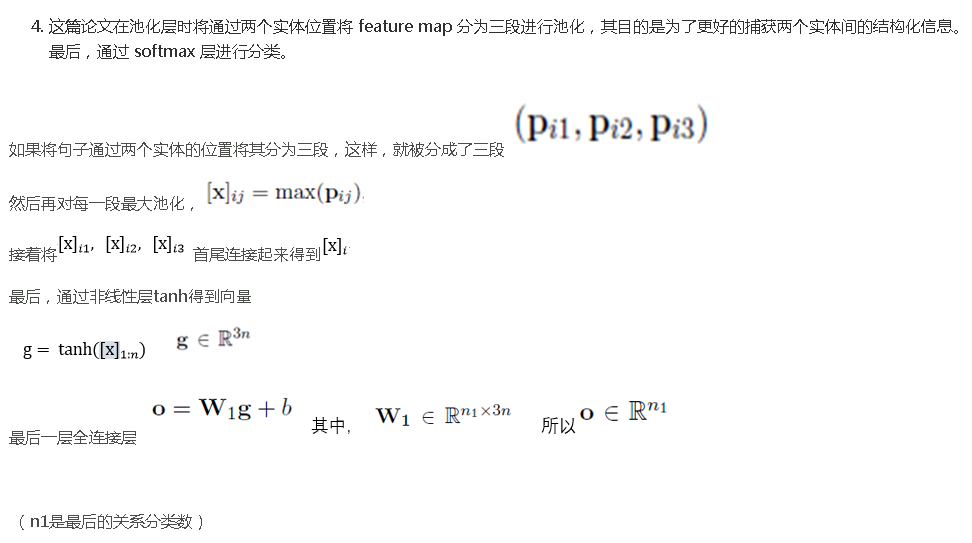


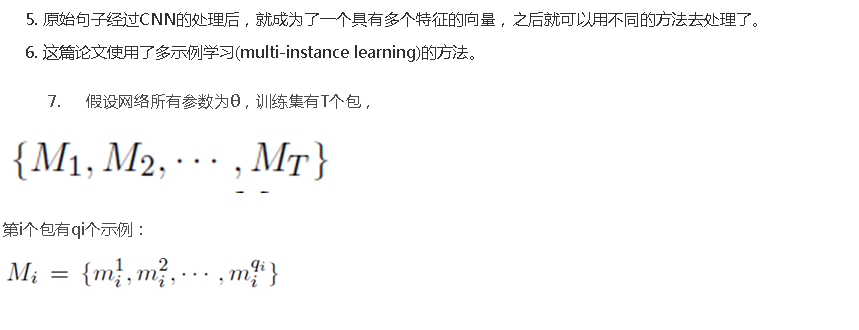
PCNN的模型比较简单，与传统卷积神经网络的区别主要在于池化层的改变。

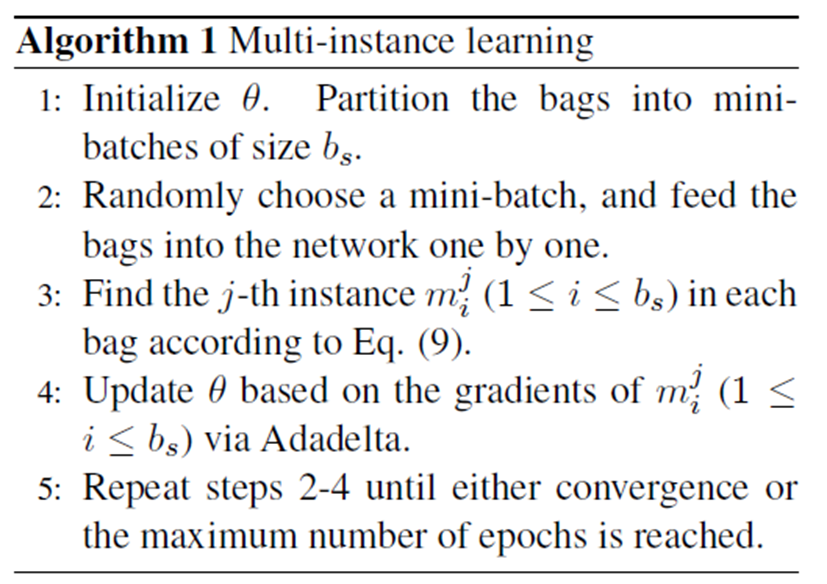
通过 word2vec 的 Skip-gram 模型将词表示成向量形式，与位置向量（各词语与两个实体的相对位置）进行拼接作为输入，之后通过卷积层得到 feature map。在池化层通过两个实体位置将 feature map 分为三段进行池化，其目的是为了更好的捕获两个实体间的结构化信息。最后，通过 softmax 层进行分类。











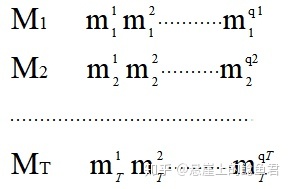
# 多示例学习（Multi-instance Learning）

多示例学习实际是一种半监督算法。考虑这样一种训练数据：我们有很多个数据包(bag)，每个数据包中有很多个示例(instance)。我们只有对bag的正负类标记，而没有对instance的正负例标记。当一个bag被标记为正时，这个包里一定有一个instance是正类，但也有可能其他instance是负类，当一个bag被标记为负类时，它里面的所有instance一定是负类。我们的目标是训练一个分类器，可以对instance的正负进行判别。

多示例学习在现实中其实很常见。如一篇文章违禁时通常是因为该文章具有某些违禁词，但我们可能无法知道具体是哪个词违禁。在这个例子中，bag就是文章，instance就是单词。

又如在医学图像领域，CT图被标定为有无病症，而一个人有病症是因为在CT图的某个区域被检测为病灶区域。我们往往只有CT图的标注，但没有CT图中每个区域的标注。这时，bag就是CT图像，而instance就是CT图中的小区域。

在这篇文章中，以句子为单位，作者构建了如下图所示的数据结构。将数据拆分成T个包，每个包包含了qT个示例。



PCNN多示例学习包结构

在数据训练的过程中，其目标函数是

