**【论文阅读笔记】 Neural Relation Extraction with Selective Attention over Instances**

<https://www.jianshu.com/p/fa2d28b437d8>

<https://blog.csdn.net/chenmingwei000/article/details/80054278>

[本文代码](https://github.com/thunlp/NRE)

Embedding(嵌入)在数学上表示一个maping, f: X -> Y， 也就是一个function，其中该函数是injective（就是我们所说的单射函数，每个Y只有唯一的X对应，反之亦然）和structure-preserving (结构保存，比如在X所属的空间上X1 < X2,那么映射后在Y所属空间上同理 Y1 < Y2)。那么对于word embedding，就是将单词word映射到另外一个空间，其中这个映射具有injective和structure-preserving的特点。

Word Embedding通俗的翻译可以认为是单词嵌入，就是把X所属空间的单词映射为到Y空间的多维向量，那么该多维向量相当于嵌入到Y所属空间中，一个萝卜一个坑。说白了，也就是找到一个映射或者函数，生成在一个新的空间上的表达，该表达就是word representation。推广开来，还有image embedding, video embedding, 都是一种将源数据映射到另外一个空间的操作。

## Abstract

* 远程监督(Distant supervised)关系抽取方法使用广泛，但不可避免地伴随错误标签问题(wrong labelling problem)，会严重影响关系抽取的效果。
* 本文提出一种基于attention的、sentence-level的方法来缓解这个问题。
* 使用CNN来嵌入句子语义，之后对多个实例(Multiple Instances)施加sentence-level的attention，来动态地减少噪声实例的权重。

## 1. Introduction

许多知识库(Freebase、DBpedia、YAGO)被用来做NLP任务，包括网络搜索和问答系统。这些知识库主要由三元组构成，例如，(Microsoft, founder, Bill Gates)。虽然包含了很多事实(facts)，但是与真实世界中存在的无限大量事实相比，还远远不够完整。

为了扩充知识库，已经有大量的工作投入在**自动寻找未知的关系事实(relation facts)** 任务中。因此，**关系抽取(Relation Extraction)** 作为从素文本中产生关系数据的过程，是NLP中一个关键性的任务。

大多数基于监督的(Supervised)关系抽取系统，需要大量带标签的、特定于关系的(relation-specific)训练数据，非常耗时耗力。Mintz在2009年提出了**distant supervisom**方法，将知识库(KBs)和文本结合起来产生训练数据。**该方法假设如果两个实体EH和ET在KB中存在某种关系A，那么所有包含这两个实体的句子都会表达同样的关系A。** 这种方法可以有效地队训练数据打标签，但是常常会产生**错误标签问题(wrong labelling problem)**，即包含EH和ET的句子中不一定表达关系A。因此，(Riedel, 2010; Hoffmann, 2011; Surdeanu, 2012)等人提出了**多样本学习(Multi-instances Learning)** 方法来缓解这个问题。

这些传统方法主要的缺点就是，大多数特征都是直接使用NLP工具来提取的，而这些方法不可避免地会传播NLP工具产生的误差。最近一些人Socher, 2012; Zeng, 2014; dos Santos, 2015)尝试使用**深度神经网络(Deep Neural Networks)** 来进行关系分类，而无需人工选择的特征。  
**特征是指的什么？要解决的问题不是如何获取大量带标签的训练数据吗？使用特征来打标签？**

这些DNN方法基于sentence-level的、带注解的数据(Annotated Data)来构造分类器。但由于缺乏人工注解的训练数据，因此难以应用到大规模的知识库(KBs)中。因此，Zeng在2015年把**多样本学习**和**神经网络模型**结合起来，根据**distant supervision**数据来构造关系提取器。**该方法假设在所有提到这两个实体的句子中，至少有一个句子表达它们的关系，并且在训练和预测过程中对每一个实体对只选择最有可能的句子。** 显然，这种方法在忽略其他句子的同时，也忽略掉了其中包含的大量信息。该方法在关系提取任务上作出了重大提升，但还不够好。  
**什么意思？？？？选择句子是为了干什么？**

* 本文提出了一种基于attention的、sentence-level的CNN模型方法来完成distant supervised关系抽取任务。
* 首先使用CNN来嵌入句子的语义(embed the semantics of sentences)。**然后就得到了sentence embeddings？** 为了利用所有的信息句，我们把关系表示成句子嵌入(sentence embeddings)的语义成分(semantic composition)。
* 此外，我们还对多个实例(Multiple Instances)施加sentence-level的attention，来动态地减少噪声实例的权重。
* 最后，我们使用由sentence-level attention赋以权重的关系向量来抽取关系。

本文的主要贡献有：

1. 与其他神经关系抽取模型相比较，该模型可以充分利用每一个实体对(Entity Pair)的所有信息句。
2. 提出了**选择性注意**(Selective Attention)方法削弱噪声实例的影响，来解决远程监督(Distant Supervision)中的错误标签问题。
3. 实验证明了Selective Attention对两种CNN模型在关系抽取任务中的表现是有益的。

## 2. Related Work

关系抽取是NLP中最重要的任务之一，学界在关系抽取方法上已经做了大量的研究，尤其是基于监督的关系抽取(Supervised Relation Extraction)。这些方法绝大多数都需要大量的带注解的数据(Annotated Data)，非常耗时耗力。

* 为了解决上述人工标注问题，Mintz在2009年使用远程监督将素文本和Freebase结合起来产生带标签的训练数据。然而，远程监督不可避免地伴随着错误标签问题。
* 为了解决上述错误标签问题，Riedel在2010年将关系抽取中的远程监督建模为一个多实例-单标签(Multi-instance Single-label)问题，(Hoffmann 2011; Surdeanu, 2012)等人在关系抽取中使用了Multi-instance Single-label学习。
  + 多实例学习(Multi-instance Learning)首先是被提出用来解决药物活性预测(Dietterich, 1997)中模糊标签(Ambiguously Labelled)训练数据的问题的。
  + 多实例学习为每一个实例考虑所有标签的可靠性，Bunescu和Mooney在2007年把弱监督和多实例学习结合起来，并将其扩展至关系抽取任务。
* 但所有上述的基于特征的(Feature-based)方法都非常依赖于由NLP工具产生的特征的质量，而且还会受误差传递问题的影响。

近几年，深度学习研究取得重大进展，很多研究者尝试使用神经网络来自动地为关系提取学习特征。Socher于2012年在关系抽取中使用了递归神经网络(RNN)。他们先将句子进行解析(Parse)，然后把语法分析树中的每一个结点表示成一个向量。此外，(Zeng, 2014; dos Santos, 2015)在关系抽取中采用了一种端到端的CNN模型。而Xie在2016年尝试加入实体的文本信息来帮助关系抽取过程。

虽然这些方法都不错，但他们还是在句子层面(Sentence-Level)进行关系抽取，并且依然没有足够的训练数据。另外，神经网络模型中用不了传统的多例学习策略。因此，Zeng在2015年将至少一个多例学习(at-least-one multi-instance learning)和神经网络结合起来在与远程监督数据上进行关系抽取。然而，他们**假设对于每一个实体对(Entity Pair)只有一个句子是活动的**。因此，他们漏掉了那些被忽略的句子中所含的大量丰富信息。本文与他们的方法不一样，我们提出了在多例上施加句子层面(Sentence-Level)的注意(Attention)，可以利用所有的信息句。  
**这么说来，Sentence Attention既可以解决错误标签问题，又可以利用所有信息句？**

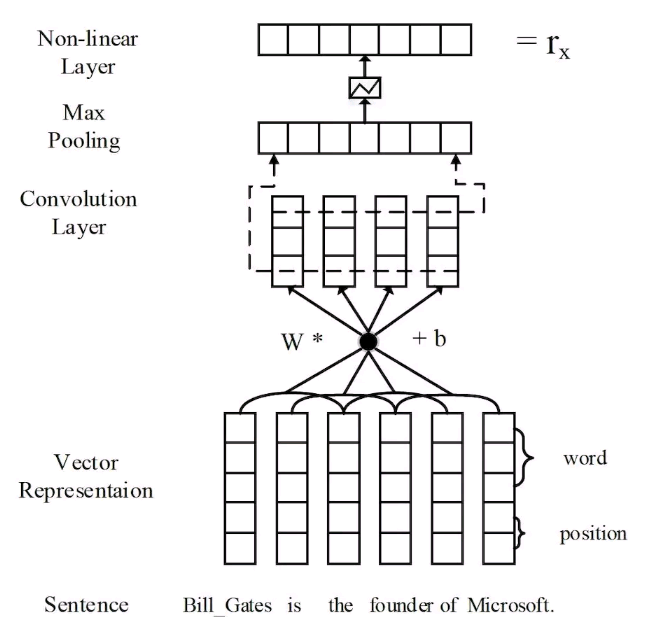
基于注意力的(Attention Based)模型最近吸引了很多研究者的兴趣，这种模型的选择性(Selectivity)使得他们能够学习不同模式(Modalities)之间的比对(Alignments。这种方法在多个领域已有应用，如图像分类、语音识别、图像字幕生成和机器翻译等。据我(Zhiyuan Liu)所知，本文是第一个在远程监督关系抽取(Distant Supervised Relation Extraction)任务中使用基于注意力(Attentiomn-Based)机制模型的研究。

## 3. Methodology

给定一系列句子{x\_1, x\_2, · · · , x\_n}和两个对应的实体，我们的模型可以测量每个关系(存在于两个给定的实体之间的)的可能性。该模型可主要分为以下两个部分：

* **Sentence Encoder**  
  给定一个句子x和两个目标实体，使用CNN来产生一个分布式的句子表示x'。(不同于句子x，x'是x的一种向量表示。)
* **Selective Attention over Instances**  
  当所有句子的分布式向量表示都得到之后，我们用句子级的注意力(Sentence-level Attention)来选择真正表达对应关系的句子。

### 3.1 Sentence Encoder



Sentence Encoder的CNN/PCNN结构

如图所示，将句子x通过CNN转换成分布式的向量表示x'。首先，句子中的词被转换成密集的实值特征向量。然后依次通过卷积层，最大池化层和非线性转换层来得到句子的分布式表示，即x'。

#### 3.1.1 Input Representation

CNN的输入是句子x中未经处理的一些词，我们首先将这些词转换成低维度的向量。本文中，每一个输入的词都会通过词嵌入矩阵(Word Embedding Matrix)来转换成一个向量。此外，本文对句子中所有的词使用位置嵌入(Position Embedding)来具体指出每一个实体的位置。

* **Word Embeddings**  
  词嵌入旨在将单词转换成分布式表示，用来获取单词的语法和语义意义。给定一个由m个词组成的句子x = ({\omega\_1,\omega\_2, · · · ,\omega\_m}), 其中的每一个词\omega\_i由一个实值向量来表示。词表示(Word Representation)由嵌入矩阵中的列向量来编码，其中嵌入矩阵是一个固定大小的词汇表。
* **Position Embeddings**  
  在关系抽取任务中，靠近目标实体的词通常能为判断实体间的关系提供有用信息。类似于(Zeng, 2014)的方法，我们使用由实体对指定的位置嵌入。它可以帮助CNN记录每个词距离头实体或尾实体多近。位置嵌入被定义为当前词与头实体/尾实体相对距离的结合。例如，在句子“Bill Gates is the founder of Microsoft.”中, “founder” 这个词到头实体Bill Gates的相对距离是3，而到尾实体Microsoft的相对距离是2。  
  **这里的Head Entity和Tail Entity应该不是指的长尾分布中的Head和Tail**

本文中假设词嵌入的维度d\_a为3，位置嵌入的维度d\_b为1(有两个位置嵌入)，可以在上图中找到此结构。最后，我们对所有的词将其词嵌入和位置嵌入组合起来，并将其记为一个向量序列。w = ({w\_1,w\_2, · · · ,w\_m}), 其中w\_i是d维向量，d = d^a + d^b \times 2。

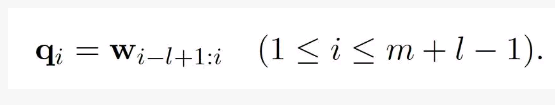
#### 3.1.2 Convolution, Max-pooling and Non-linear Layers

在关系抽取中，主要的挑战是**句子的长度具有可变性**。此外，句子中重要的信息可能出现在一句话中的**任意位置**。因此，我们应该利用所有的局部特征，并在全局范围内进行关系预测。本文中使用一个卷积层来合并这些局部特征。首先，卷积层使用长度为L的滑动窗口在句子上滑动提取局部特征。之前提到过，本文中假设词嵌入的维度为3，因此这里设置的窗口大小L为3。然后将所有的局部特征结合，通过最大池化操作来得到一个固定大小的向量。这样就将长度不定的句子输入转换成了一个固定长度的向量表示，换句话说，就是将一个sentence转换成了一个sentence embedding。  
**这个所谓的固定长度是怎么计算的？？？**

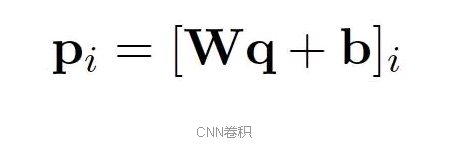
**下面就是讲这个利用卷积神经网络(CNN)转换的具体过程。**

本文中，卷积被定义为一个向量序列w和卷积矩阵W之间的操作。**向量序列w**是有m个(句子中词的个数)向量的序列，其中每个向量的维度为d。 **词嵌入矩阵W**的维度为W \in R^{d^c \times L \times d}，其中d^c是句子嵌入的大小。定义向量q\_i \in R^{l \times d}为在第i个窗口里的连续的多个词\omega的词嵌入(w)的级联(一系列互相关联的事物)。

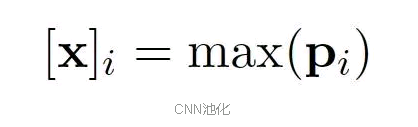
注意\omega表示的是句子中的词(word)，而w表示的是词(word)对应的词嵌入(word embedding)。其实都是一个东西，不用分太清楚。。打字的时候也不太好区分，可能有些地方搞混了。只要明白有这么个映射(embedding)过程就行了。



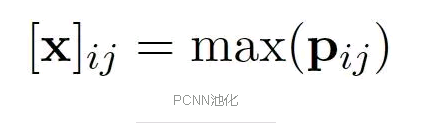
由于窗口在边界滑动时可能会超出句子的边界，我们为每个句子设置了特殊的填充。即将所有超出边界的输入向量\omega\_i(i \lt 1 or i \gt m)当做0向量。因此，卷积层的第i个卷积核(Filter, namely kernal)，可按如下方式计算：



其中b是偏置向量，**W** 是卷积矩阵, q是上式定义的级联。向量x \in R^{d^c}的第i个分量如下：



进一步的，Zeng在2015年提出了PCNN(CNN的一种变体)，在关系抽取中使用了分段的(piecewise)最大池化操作。每一个卷积核p\_i被头实体和尾实体分为三段(p\_{i1}, p\_{i2}, p\_{i3})，相应地，最大池化过程也分为三个部分独立进行。



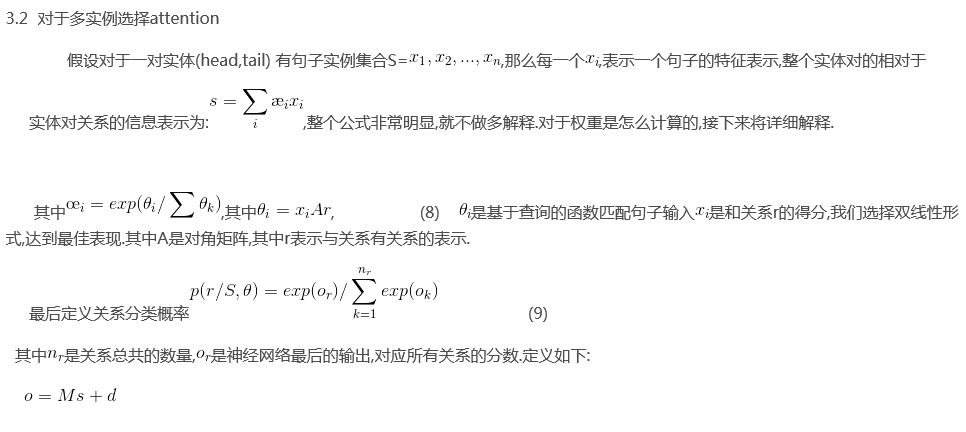
[x]\_i就是[x]\_{ij}的级联集合。

最后在输出层使用非线性激活函数，例如双曲正切函数(tanh)。  
**到最后还是没说句子嵌入(Sentence Embedding)的Size怎么计算的啊？$d^c$是怎么计算的？**

### 3.2 Selective Attention over Instances

假设一个集合S中含有与实体对 (head, tail) 相关的n个句子，S = {x\_1, x\_2, ..., x\_n}。为了利用所有句子中的信息，将集合S表示成一个向量s，显然，该向量s取决于集合S中所有句子 x\_i的向量表示*x\_i*。每一个句子的向量表示都包含关于在该句子x\_i中实体对 (head, tail) 是否包含关系r的信息。向量 s 是集合S中所有句子的向量表示的加权和。  
**这个关系r指的是某一个特定的关系？那岂不是要对所有的关系都进行一次计算？或者r指的是所有关系的一个综合性的向量表示？**

* Baseline  
  认为每个句子都具有同等贡献，因此权重都为1/n。
* Selective Attention  
  为每个句子赋以不同的权重，根据输入句x\_i与关系r的匹配程度赋值。
* 此外，还提出了一种计算可能性的方法。o表示神经网络的输出，维度与所有关系类型的个数相同，o\_i对应的数值表示在句子集S中实体对 (head, tail) 与关系r\_i的匹配程度。



Zeng在2015年提出假设，在所有的实体对提及 (entity pair mention, 所谓实体对提及，也就是一类句子，这类句子中包含了考虑的两个实体head和tail提及)中至少有一个提及能够表示他们之间的关系，并且只用句子集中最可能的那一个句子来训练。

### 3.3 Optimization and Implementation Details

* 交叉熵作为优化目标函数，使其最小
* SGD随机梯度下降算法
* 使用Dropout策略避免过拟合

## 4. Experiments

实验是为了证明提出的句子层面的Selective Attention能够有效缓解distant supervised关系提取方法中的错误标签问题，并有效利用所有的信息句。

### 4.1 Dataset and Evaluation Metrics

* 使用常用[数据集](http://iesl.cs.umass.edu/riedel/ecml/)来进行评测。
* Entity Mentions是用Stanford命名实体tagger得到的。
* 使用held-out评测方法，与Freebase进行比较。

### 4.2 Experimental Settings

#### 4.2.1 Word Embeddings

* 使用[word2vec工具](https://code.google.com/p/word2vec/)在NUT文集上训练词嵌入
* 把在文集中出现过100次以上的词当做词汇表
* 当一个实体有多个单词时，就把这些单词连起来

#### 4.2.2 Parameter Settings

* 在训练集上使用三折交叉检验来调整模型

### 4.3 Effect of Sentence-level Selective Attention

* 加上ATT的CNN和PCNN都比之前的效果好，说明ATT(Sentence-level Attention)是有效的

### 4.4 Effect of Sentence Number

* 略

### 4.5 Comparison with Feature-based Approaches

* 略

### 4.6 Case Study

* 略

## 5. Conclusion and Future Works

* 我们把CNN和Sentence-level的Selective Attention结合起来了。既有效利用了所有信息句的信息，又缓解了distant supervised关系抽取方法中的错误标签问题。而且效果比以前的方法都好。
* 未来计划把该方法应用在其他任务上，比如说text categorization。
* CNN只是众多神经关系抽取方法中的一个模型，还有其他网络模型可以拿来和ATT融合。