Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

关系抽取之远程监督

白崇有　硕士

2019年08月18日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

背景简介

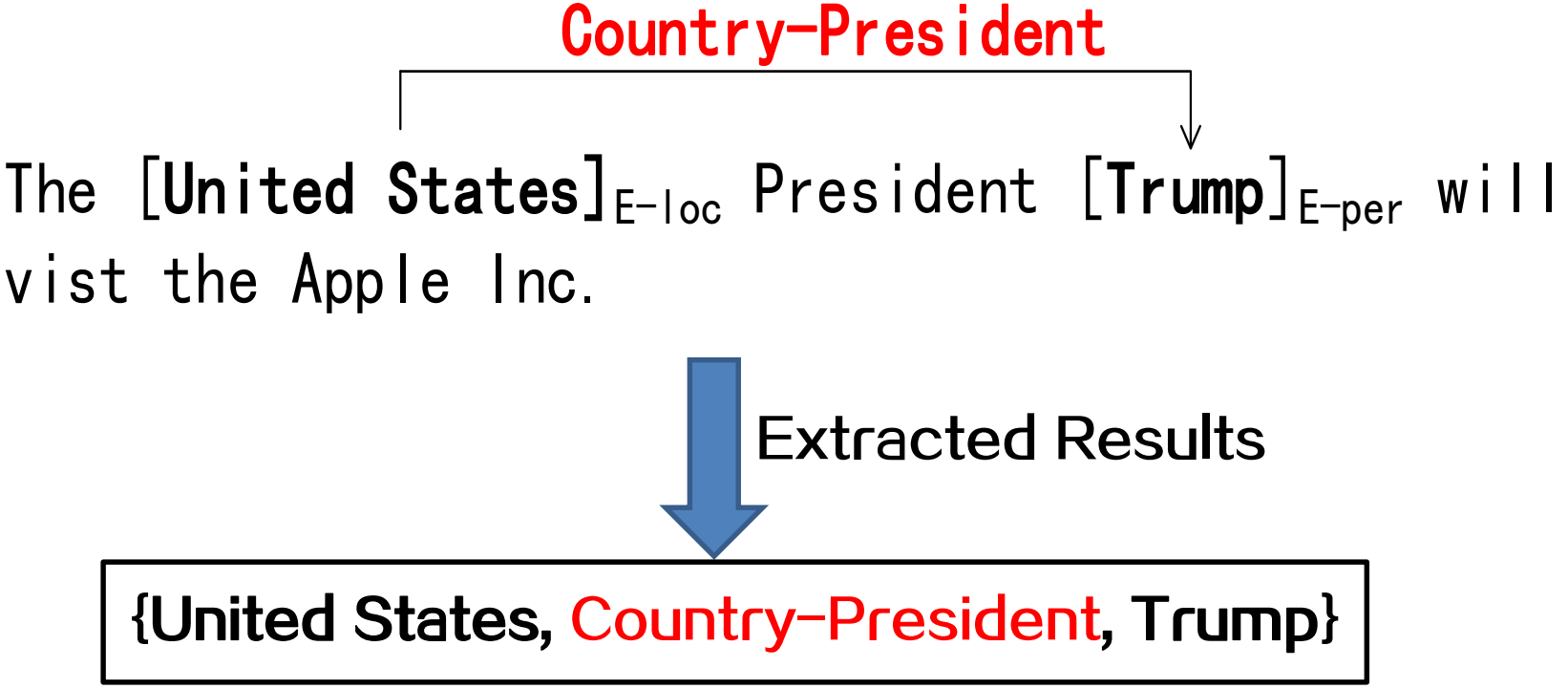
- 预期收获
  - 1. 了解远程监督
  - 2. 了解远程监督应用到关系抽取任务会遇到的问题
  - 3. 了解关系抽取中两种去除噪声样本的方法

背景简介

- 远程监督方法提出的背景
  - 人工标注训练语料费时费力，而且数量有限，数据规模大大限制了模型训练。
  - 2009年，Mintz首次将远程监督方法应用到了关系抽取任务。

基本概念

- 关系抽取
  - 位于命名实体识别任务之后，从文本中抽取实体对之间的语义关系。



- 基本概念

关系抽取的方法

1.无监督的学习方法

2.半监督的学习方法

3.有监督的学习方法

- 3.1实体和关系的联合学习方法
- 3.2只抽取关系的学习方法

4.远程监督的学习方法

基本概念

- 基本概念
  - 在2009年，Mintz首先使用远程监督的方法来构造训练数据集，并将得到的数据集应用到关系抽取任务上。
  - 远程监督（Distant Supervision）：远程监督就是将已有的知识库（比如 freebase）对应到非结构化数据中（比如新闻文本），从而生成大量的训练数据。

基本概念

- 远程监督的假设

  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.（如果训练语料中的句子所包含的实体对在知识库中有关系的体现，那么我们认为训练语料中所有包含相同实体对的句子都可以表达此关系类型。）

基本概念

Freebase

Relation	Entity1	Entity2
founders	Apple	Steve Jobs
…	…	…

free texts

1.**Steve Jobs** was the co-founder and CEO of **Apple** and formerly Pixar.
2.**Steve Jobs** passed away the day before **Apple** unveiled iPhone 4s in late 2011.

基本概念

- 远程监督应用到关系抽取任务上存在的问题
  - 远程监督的假设过于肯定，按照这个假设去构建训练数据会带来大量的噪声样本。
  - 知识库不完善，也会引入噪声样本

基本概念

- 基本概念
  - 多示例学习

    多示例学习可以被描述为：假设训练数据集中的每个数据是一个包(Bag)，每个包都是一个示例(instance)的集合,每个包都有一个标签，而包中的示例是没有标签的；如果包中至少存在一个正标记的示例，则包被赋予正标签；而对于一个有负标记的包，其中所有的示例均为负标签。
  - 《多示例学习-白崇有》

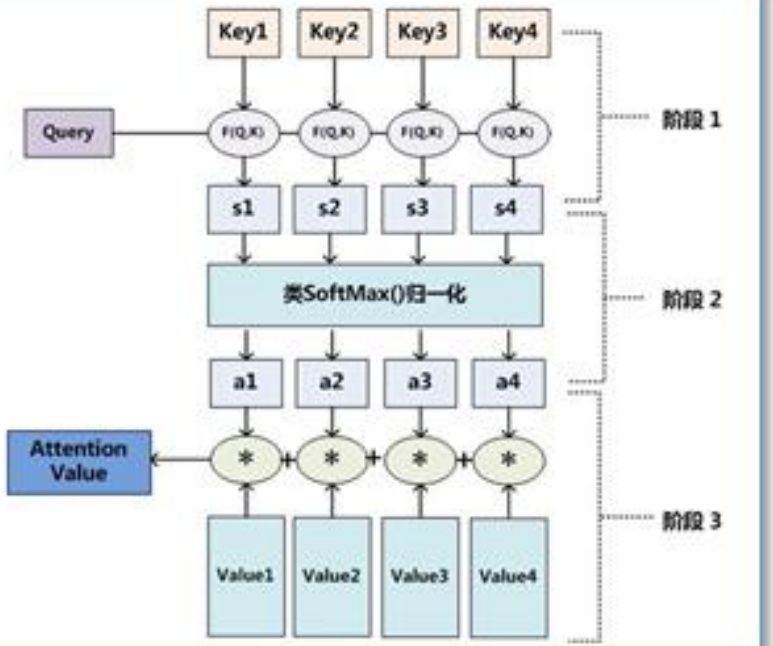
基本概念

- 基本概念
  - 注意力机制

$$f(Q,K_i)=\begin{cases} Q^T K_i & dot \\ Q^T W_a K_i & general \\ W_a[Q;K_i] & concat \\ v_a^T \tanh(W_a Q + U_a K_i) & perceptron \end{cases}$$

$$a_i = soft\max(f(Q,K_i)) = \frac{\exp(f(Q,K_i))}{\sum_j \exp(f(Q,K_j))}$$

$$\text{Attention}(Q,K,V) = \sum_i a_i V_i$$



  - 《2018.10.07-注意力机制-白崇有》

算法原理

T	使用远程监督方法产生的训练集训练得到一个高效的关系分类器
I	通过远程监督方法产生的Bag集合
P	1.嵌入；2.句子编码 3.句子选择；4.Bag向量表示 5.全连接层；6.Softmax
O	关系分类器

P	通过远程监督方法构造的训练集中会存在噪声样本
C	远程监督方法是有效的，可以为训练语料中的大部分句子标记正确的标签
D	如何去除通过远程监督方法构造的训练集中的噪声样本?
L	ACL

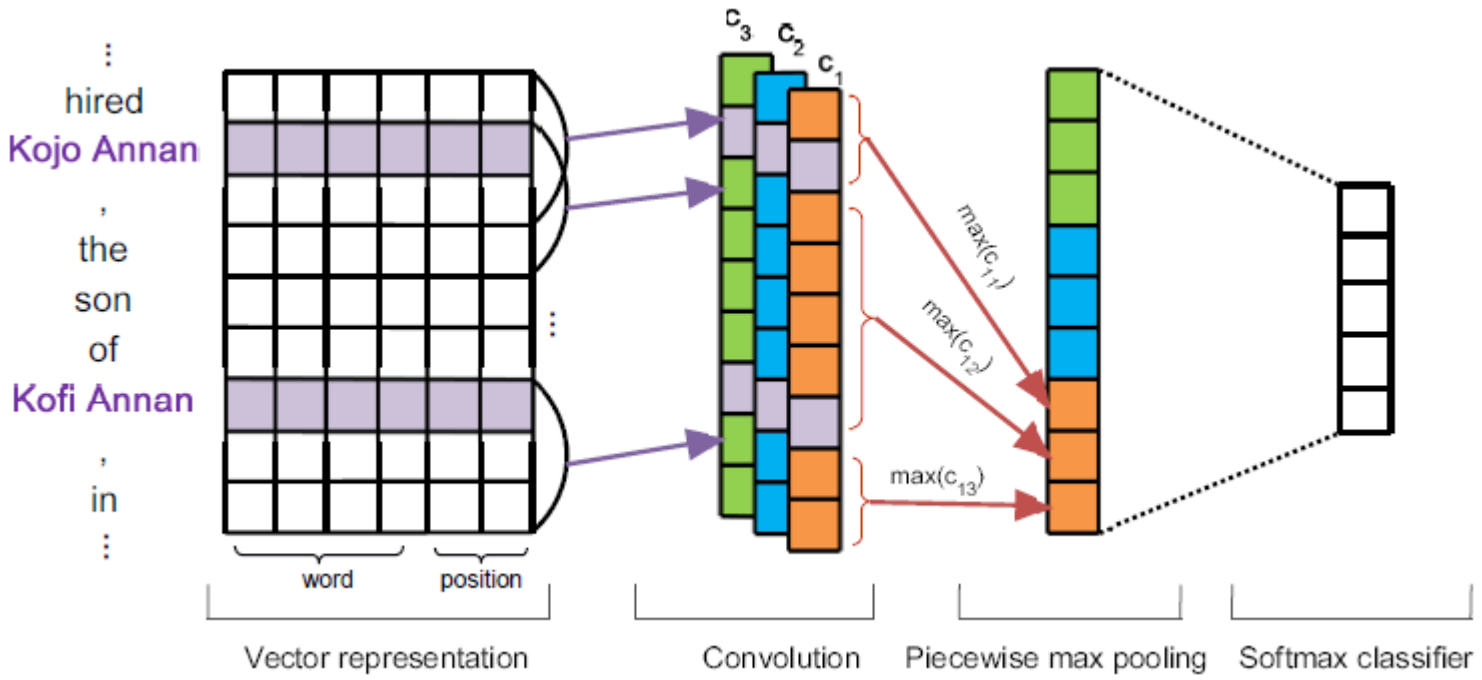
算法原理

- 算法流程
  - 远程监督关系抽取被视为多示例学习问题

多示例学习	
1:	将训练集中所有的句子都打包成Bag
2:	初始化参数θ，把bag分成大小为b的mini-batches
3:	随机选择一个mini-batch的bag，然后输入到网络中
4:	对bag中的所有句子分别进行编码
5:	按照一定的示例选择方法缓解bag中噪声样本的影响
6:	更新参数θ
7:	重复步骤3-6，直到模型收敛或者达到最大的迭代次数

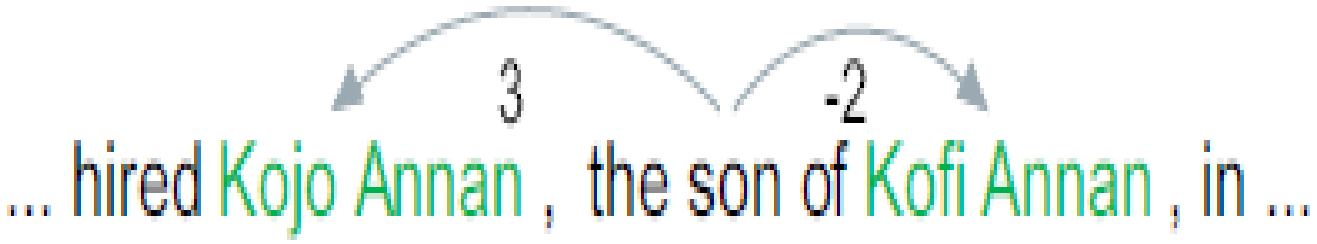
算法原理

- 句子编码器–PCNN



算法原理

- Position Embeddings
  - 使用位置特征（PFs）去识别实体对
  - 一个PF被定义为当前词到实体1和实体2之间相对位置的组合



  - 随机初始化两个位置嵌入矩阵（PF1和PF2），通过查找这两个位置嵌入矩阵将相对位置转为实值向量，最后将得到的这两个位置的向量进行拼接得到position embeddings。

算法原理

- PCNN
  - PCNN与CNN的比较
  - ➢ 卷积过程相同
  - ➢ 池化过程不同

在关系抽取中，一个输入的句子根据实体的位置将句子分成3个片段，PCNN分别对每个片段进行池化，将这三段池化的结果进行拼接。

  - PCNN的输出

$$g = \tanh(c)$$

其中，c是PCNN池化之后产生的一个句子向量，然后再通过一个非线性函数（tanh）进行激活。

算法原理

- 如何选取bag中的句子?
  - 第一种方法

    基于多示例学习的至少一个假设，选择bag中置信度最高的一个句子用于训练

算法原理

- 如何选取bag中置信度最高的句子?

$$o = W \cdot g_i^j + b$$

其中，W是一个矩阵，g_i^j表示第i个bag中的第j个句子经过PCNN得到的句子向量，o是一个长度为n的向量，n表示可能的关系类型的数量，o^r 表示第r种关系的分数。

$$p(r|x_i^j;\theta) = \frac{e^{o^r}}{\sum_{k=1}^{n} e^{o^k}}$$

其中，x_i^j表示第i个bag中的第j个句子，$p(r|x;\theta)$表示条件概率分布

算法原理

- 如何选取bag中置信度最高的句子?

$$j^* = \underset{j}{\operatorname{argmax}}\, p\left(y_i \middle| x_i^j; \theta\right) \quad 1 \le j \le q_i$$

其中，y_i是目标输出，j^*表示选取第j^*的句子， q_i表示第i个bag中句子的数量。

- 目标函数

$$J(\theta)=\sum_{i=1}^{T} \log p(y_i|x_i^{j^*};\theta)$$

其中，T表示训练集中bag的数量

算法原理

- 第一种方法存在的问题
  - 1. 选取每个句子bag中置信度最高的样例作为正样例进行训练，在滤除噪声的同时也丢失了很多有用的监督信息
  - 2.忽略了bag内的所有句子都是噪声样本这种情况

算法原理

- 如何选取bag中的句子?
  - 第二种方法

    利用Attention机制来降低bag中噪声样本的权重，即可以充分利用到bag中所有标记正确的句子又可以缓解bag中噪声样本的影响。

算法原理

- 如何选取bag中的句子?
  - 给bag中的每个句子分配一个权重，然后通过加权求和得到bag水平的向量表示s, 用s来预测关系类型

$$s = \sum_{i=1}^{q} \alpha_i \cdot g_i$$

其中，q为bag中句子的数量，g_i 是第i个句子经过PCNN编码产生的向量，α_i 表示bag中第i个句子的权重。

算法原理

- 如何得到句子的权重?
  - 平均分配策略
    - 假设bag中的每个句子对bag具有相同的贡献

$$\alpha_i = \frac{1}{q}$$

算法原理

- 如何得到句子的权重?
  - Attention机制
  - 1. 相似度计算

$$e_i = g_i A r$$

其中，e_i表示 g_i 和目标关系r的匹配程度，r表示关系r的向量表示。

  - 2. 权值归一化

$$\alpha_i = \frac{\exp(e_i)}{\sum_k^q \exp(e_k)}$$

算法原理

- 如何得到句子的权重?
  - 3. 加权求和得到bag水平的向量表示s

$$s = \sum_{i=1}^{q} \alpha_i \cdot g_i$$

算法原理

- 通过bag的向量表示s来预测关系?

$$o = Ms + d$$

其中，o是一个长度为n的向量，n表示可能的关系类型的数量，o^r 表示第r种关系的分数。

$$p(r|s;\theta) = \frac{e^{o^r}}{\sum_{k=1}^{n} e^{o^k}}$$

其中，s表示bag的向量表示，$p(r|s;\theta)$表示条件概率分布

- 目标函数

$$J(\theta) = \sum_{i=1}^{T} \log p(y_i|s;\theta)$$

其中，T表示训练集中bag的数量

算法原理

- 第二种方法存在的问题
  - 1.忽略了bag内的所有句子都是噪声样本这种情况

优劣分析

- 优势
  - 方法1和方法2实现比较简单
- 劣势
  - 忽略了bag内的所有句子都是噪声样本这种情况

应用总结

- 远程监督的应用领域
  - 命名实体识别
  - 关系抽取
  - …
- 基于远程监督的关系抽取未来的研究方向?
  - 1.在句子水平上去噪
  - 2.在bag水平去噪，主要针对的是bag中所有句子都是噪声样本的这种情况
  - 3.充分利用知识库中的一些监督信息来缓解远程监督噪声的影响
  - 4.…

参考文献

[1] Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1753-1762.

[2] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 2124-2133.

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！