Distant Supervision via Prototype-Based Global Representation Learning

huweilong@whu.edu.cn

武汉大学计算机学院

2017 年 9 月 28 日

Table of contents

1 简介
- 背景
- 主要工作

2 相关工作

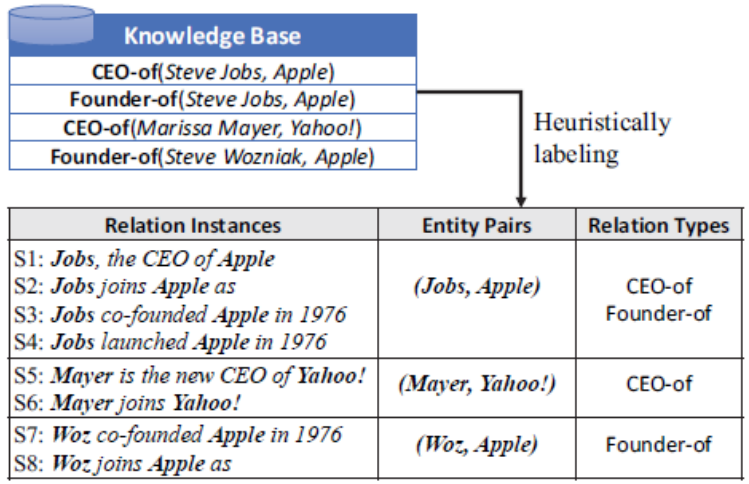
3 文章方法
- 基于原型的实体对嵌入
- 原型学习
- 神经网络

4 实验
- 数据集和 Baselines
- 结果与分析

5 结论

1 传统关系抽取方法缺点：

- 有监督的
- 在 web 数据或者开放领域中存在标记数据短板



2 为了解决上述问题，DS 被提出

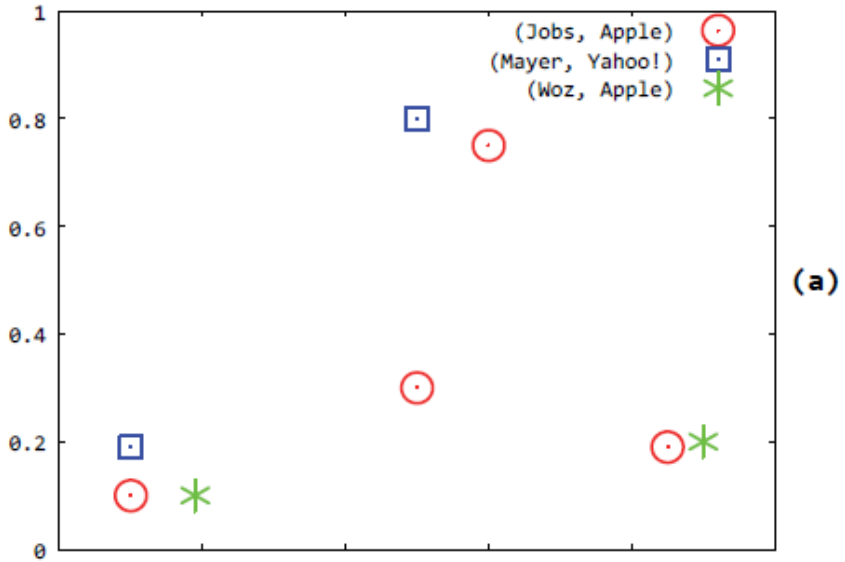
3 DS 面临的问题：

- multi-instances problem：对实体对分类，每对实体包含诸多实例
- missing-lable problem：只有实体的标签给定

instance-level models 先学习实例级别的分类器，每对实体的关系类型取决于它所有实例的分类结果。

instance-level models 中的两个问题:

- 无法区分对判别实体对关系类型无关联的信息
  例如，$Founder-of(Jobs, Apple)$：S3、S4 是相关的，S1、S2 是无关的
- 难以学习到准确的分类器



学习实体对在全局特征空间中的表示

本文的三点改进

1. 解决 multi-instance 和 missing label 问题
2. 捕获多个实例信息，在实体对识别易于分类
3. 过滤不相关的实例

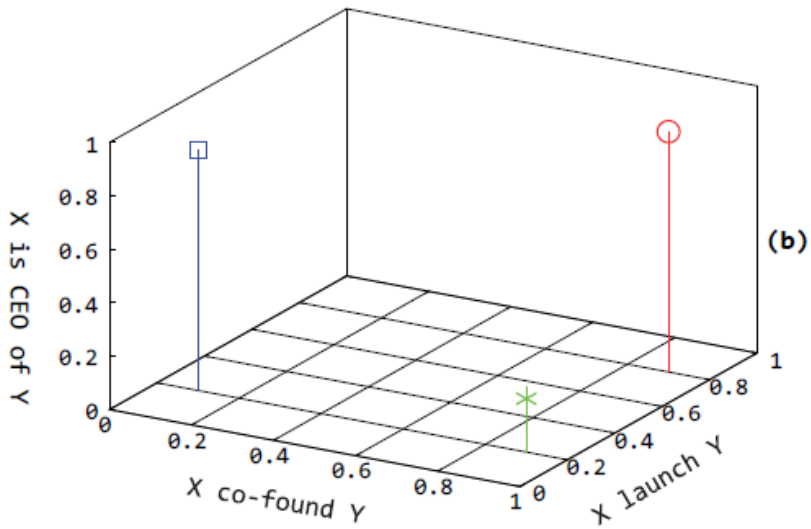


Table of contents

1 简介
- 背景
- 主要工作

2 相关工作

3 文章方法
- 基于原型的实体对嵌入
- 原型学习
- 神经网络

4 实验
- 数据集和 Baselines
- 结果与分析

5 结论

两种通用策略：

1. multi-instance learning techniques
   - 通过实例标签和实体对标签建模关系标签，从而学习实例级别的分类器
   - at-least-one assumption, relational classifier, Markov Logic Network ……
   - 因为 missing label 问题，multi-instance 模型的学习相当困难
2. better training instance labeling algorithms
   - 简单的 DS 假设导致错标训练样本
   - 因此很多算法注重消除错误标记的样本实例

其它策略：

- 在标记语料库中加入启发式标记的 DS 语料
- 运用 relations/instances/features 的同现统计
- 运用分段 CNN 表示关系实例

Table of contents

1 简介
- 背景
- 主要工作

2 相关工作

3 文章方法
- 基于原型的实体对嵌入
- 原型学习
- 神经网络

4 实验
- 数据集和 Baselines
- 结果与分析

5 结论

- 目标：

  给定实体对 **B**，从它的所有实例中寻找包含所有相关信息的全局特征向量

- 假设：

  每种关系类型存在一系列原型，利用这些原型能够推断出该关系

  Founder-of："X is the founder of Y", "X co-found Y" and "X launch Y in....."

  "Jobs co-founds Apple in 1976" – "X co-found Y"

如果实体对包含大量与某个关系的原型相近的实例，该实体对的很有可能表达该关系，正式地：

原型：$C = c_1, c_2, ..., c_k,$　　则 B 的特征向量为

$m(B) = [m_1(B), m_2(B), ..., m_k(B)]$, 其中 $sim(x_i, c_k, w_k) = \sum_j w_{kj} x_{ij} c_{kj}$ 等

$$m_k(B) = \max_i sim(x_i, c_k, w_k)$$

$$\begin{bmatrix} max\ sim(., X\ co-found\ Y) \\ max\ sim(., X\ launch\ Y) \\ max\ sim(., X\ is\ CEO\ of\ Y) \end{bmatrix} = \begin{bmatrix} sim(S3, X\ co-found\ Y) \\ sim(S4, X\ launch\ Y) \\ sim(S1, X\ is\ CEO\ of\ Y) \end{bmatrix} = \begin{bmatrix} 0.9 \\ 0.8 \\ 0.9 \end{bmatrix}$$

结论：

1. 全局表示方式能从多个实例中综合相关的信息
2. 全局表示方式能过滤无关的实例信息，即过滤噪音数据

原型必须具备两个要求：

- *Goodness-of-exemplar* 可以概括关系类型所有实例的中心趋势
- *Goodness-of-discrimination* 对关系的区分度高

给定特定关系的训练实例 $x_1, x_2, \ldots, x_m$，训练步骤如下：

1. 利用算法初始化原型集合
2. 利用现有原型训练关系抽取模型
3. 根据错分的实体对收集错分的训练实例
4. 利用算法从错分的训练实例中重新采样新的原型，加入到原型集合
5. 回到第二步，重复直到收敛

Weighted Rejection Sampling Algorithm
Input: - The wrongly classified instances $\mathbf{X} = \{\mathbf{x}_1, \ldots \mathbf{x}_m\}$ - The number of sampled prototypes K - The similarity threshold σ **Output:** The new prototypes $\mathbf{C}=\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$

For $\mathbf{x}_i$ in $\mathbf{X}$:
Compute σ-NN($\mathbf{x}_i$)
End for
$\mathbf{C} \leftarrow \{\}$
While Size($\mathbf{C}$) < K:
Sample $\mathbf{x}$ from $\mathbf{X}$ with probability $\propto \exp(\sigma\text{-NN}(\mathbf{x}))$
If $\max_k \text{sim}(\mathbf{x}, \mathbf{c}_k) < \sigma$:
Add $\mathbf{x}$ to $\mathbf{C}$
End while

- prototype-base embedding

  prototype similarity layer and max-pooling layer
- entity pair classification

  linear SVM layer
- learn parameters

  AdaDelta optimization algorithm
- solve multi-label problem

  each relation type a model using the "one-versus-all" strategy

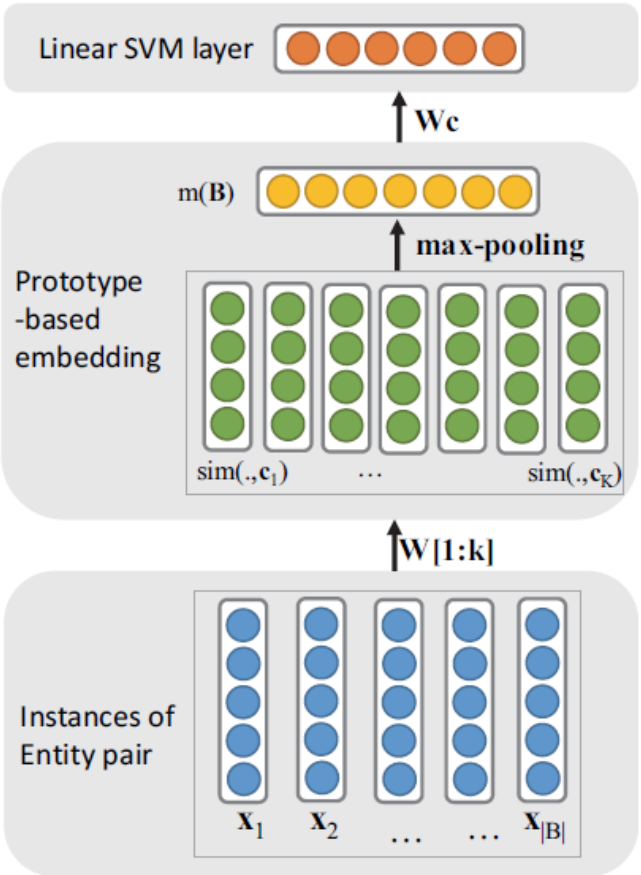


Table of contents

1 简介
- 背景
- 主要工作

2 相关工作

3 文章方法
- 基于原型的实体对嵌入
- 原型学习
- 神经网络

4 实验
- 数据集和 Baselines
- 结果与分析

5 结论

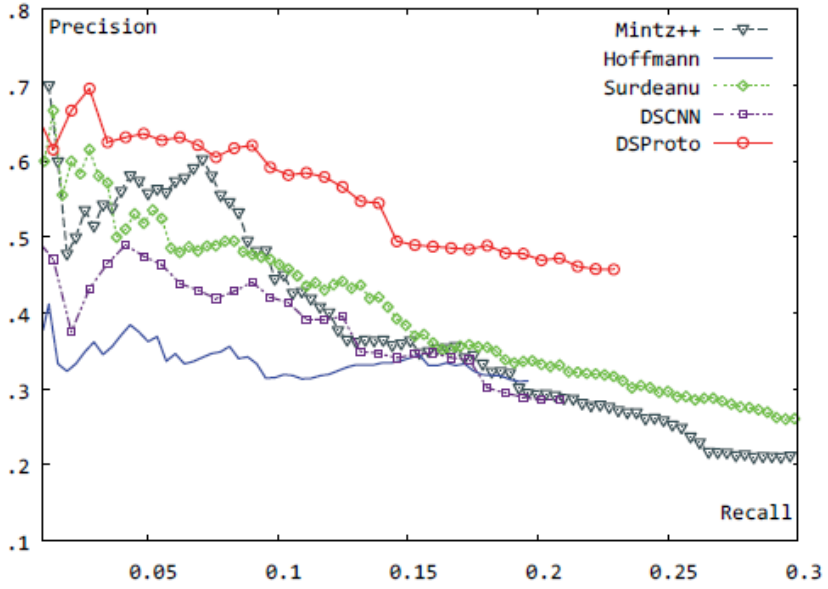
KBP 数据集 (Surdeanu et al., 2012)：

- 41 种关系类型，183,062 个训练实例和 3,334 个测试实例
- 评价关系文档无关、只考虑句中主要关系

Baselines:

- *Mintz++* 用所有实例特征表示实体对
- *Hoffmann* at-least-one assumption
- *Surdeanu* relational classifier
- *DSCNN* Convolutional layer + Max-pooling layer

结果：



System	P	R	F1
Mintz++	0.260	0.250	0.255
Hoffmann	0.306	0.198	0.241
Surdeanu	0.249	0.314	0.278
DSCNN	0.286	0.214	0.244
DSProto	**0.459**	**0.231**	**0.307**

分析：

The effect of the size of prototypes

	25%	50%	100%	200%
KBP	0.272	0.283	**0.307**	0.298

The effect of iterative prototype learning

	One-Shot	Iterative
KBP	0.286	**0.307**

Table of contents

1 简介
- 背景
- 主要工作

2 相关工作

3 文章方法
- 基于原型的实体对嵌入
- 原型学习
- 神经网络

4 实验
- 数据集和 Baselines
- 结果与分析

5 结论

This paper describes a new distant supervision paradigm - global repersentation learning-based distant supervison and proposes an effective global representation learning algorithm - prototype-based embedding. By learning informative entity pair representations, our method can achieve competitive performance. This paper uses manually designed instance features to represent instances, in future wu want to develop a neural network which can jointly embed relation instances and entity pairs.

谢谢！

huweilong@whu.edu.cn