
知识图谱
技术分享
Multipurpose Presentation

目录

C O N T E N T S

3.5、实体连接

- 候选实体生成：
- 根据输入文本中检测出的实体mention集合M，从给定知识图谱中找到每个实体m属于M可能对应的候选实体集合Em。

- 候选实体排序：
- 负责对候选实体集合Em中多个候选实体打分的排序，并输出得分最高的候选实体作为m的实体链接结果。

4 NRE
关系抽取（分类）

4.1、关系分类

– 任务介绍
– 数据集
– 标注工具

基于规则的方法
– 基于人工模板的方法
– 基于统计模板的方法

本章目录

基于监督的方法
– CNN
– RNN
– PCNN

半监督的方法
– 自举
– 远程监督

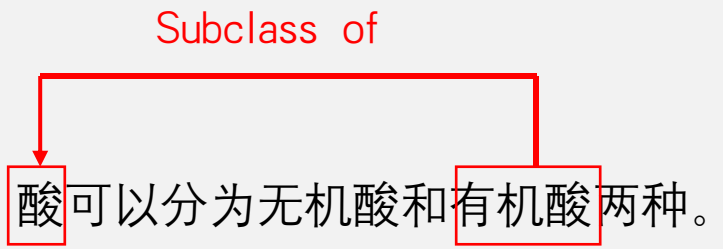
➢ https://blog.csdn.net/imsuhxz/article/details/83780454

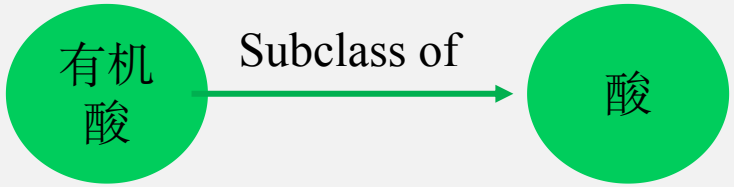
4.1、关系分类——任务简介

关系抽取：从一个句子中判断两个entity是否有关系，一般是一个二分类问题，指定某种关系

关系分类：一般是判断一个句子中两个entity是哪种关系，属于多分类问题。

- 关系提取是从语句中获取实体之间关系的一种技术
- 实体对应知识图谱中的结点，关系则对应知识图谱中的边





4.1、关系分类——数据集

1、ACE 2005：599 docs. 7 types；

2、SemiEval 2010 Task8 Dataset：

19 types

train data：8000 test data：2717

3、NYT+FreeBase 通过Distant Supervised method 提取，里面会有噪音数据：

53 types

train data：522611 sentences； test data：172448 sentences；

The <e1>microphone</e1> converts sound into an electrical <e2>signal</e2>.

Cause-Effect(e1,e2)

m.0ccvx m.05gf08 queens belle_harbor /location/location/contains .....officials yesterday to reopen their investigation into the fatal crash of a passenger jet in belle_harbor , queens.....
###END###

4.2、关系分类——标注工具

BRAT是一个基于web的文本标注工具，主要用于对文本的结构化标注，用BRAT生成的标注结果能够把无结构化的原始文本结构化，供计算机处理。利用该工具可以方便的获得各项NLP任务需要的标注语料。

4.2、基于规则的方法——人工模板

- 其中每个NP_i(i>=1)和NP_0之间都满足上下位关系

（1）	NP_0 such as $\{NP_i,\}$ * $\{(or\|and)\}$ NP_n
输入例子	. . . animals such as cat . . .
输出例子	<cat，Hyponymy，animal>
（2）	such NP_0 as $\{NP_i,\}$ * $\{(or\|and)\}$ NP_n
输入例子	. . . such authors as Herrick and Shakespeare . . .
输出例子	<Herrick，Hyponymy，author>；<Shakespeare，Hyponymy，author>

- badcase: animals other than dogs such as cats
- animal和dogs/cats

4.2、基于规则的方法——基于统计的方法

Ravichandran等人提出了，基于搜索引擎的统计模板抽取方法，抽取结果可以用于关系分类任务和答案抽取任务。

首先，从全部待分类关系中选择一个关系，例如birthday，并找到满足该关系的一个实体对，例如mozat（对应问题实体<name>）和1755（对应答案实体<answer>）。

然后将该实体对作为查询语句，例如Mozart+1756，提交到搜索引擎，并抓取搜索引擎返回的前n个结果文档。接下来，保留返回结果文档中同时包含该实体对的句子集合，例如：(a)The great composer Mozart(1756–1791) achieved fame at a young age.(b)Mozart(1756–91)was a genius.(c)The whole world would always be indebted to the great music of Mozart(1756–1791)，并对每个句子分词。

最后，从保留句子的合集中寻找包含上述实体对的最长字串，例如Mozart（1756–，并将实体替换为非终结符得到一个模板，例如<NAME>(<AVSWER>–。同一个关系使用不同的实体对能够抽取得到不同的模板。例如，关系Birthday抽取的模板候选包括：（a）<Name>(<ANSWER>–、（b）born in <ANSWER>, <NAME>和（c）<NAME> was born on <ANSWER>等。

4.2、基于规则的方法——基于统计的方法

那么如何计算自动抽取出来的模板的置信度呢？

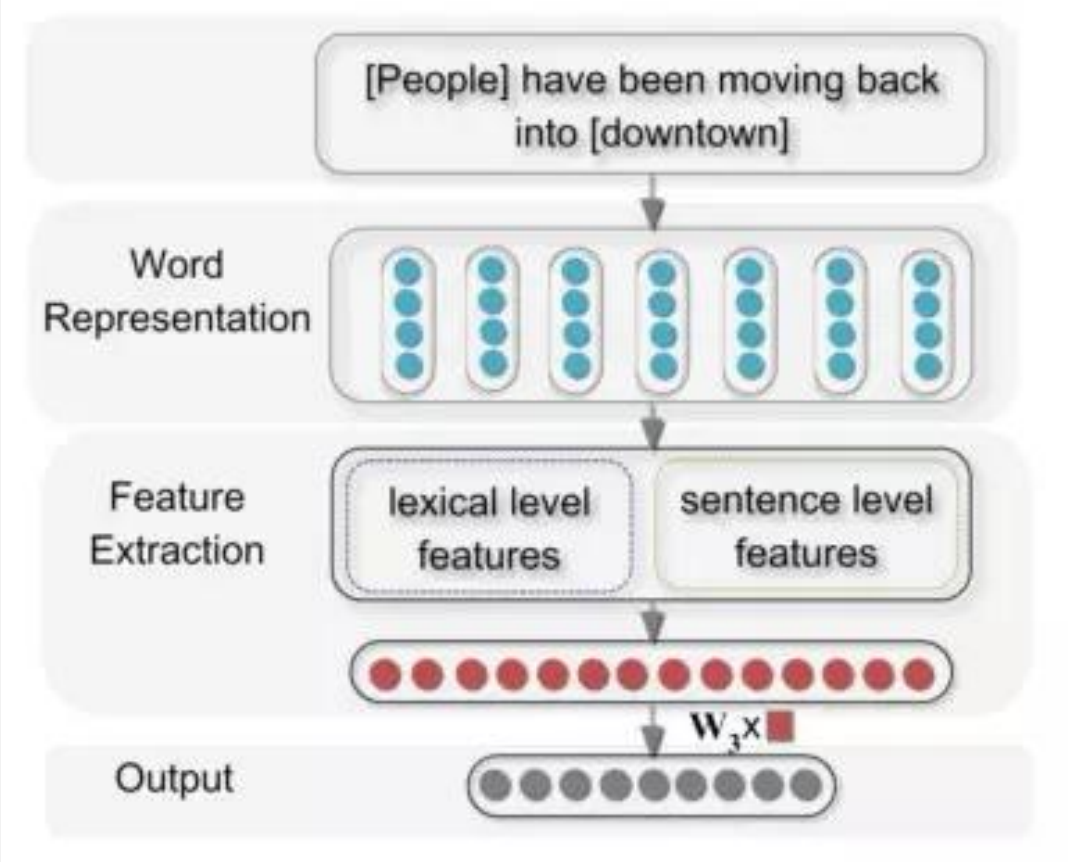
给定一个关系，首先，选择满足当前关系的一个实体对（例如Mozart和1756），将该实体对中的问题实体（例如Mozart）单独作为查询语句提交给搜索引擎，并保留返回结果文档中包含该问题实体的全部句子。

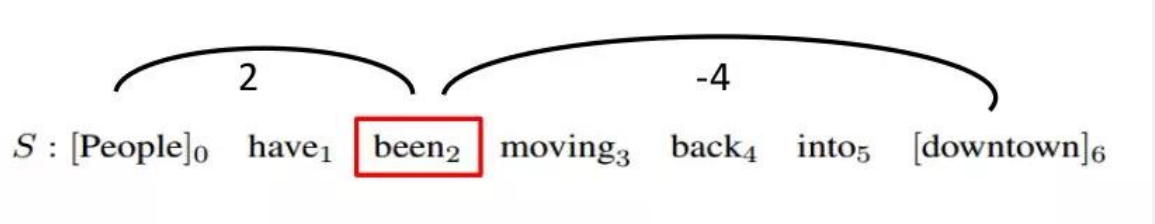
然后，计算给定模板$pattern_i$在该句子集合上的对应得分P（$pattern_i$）：

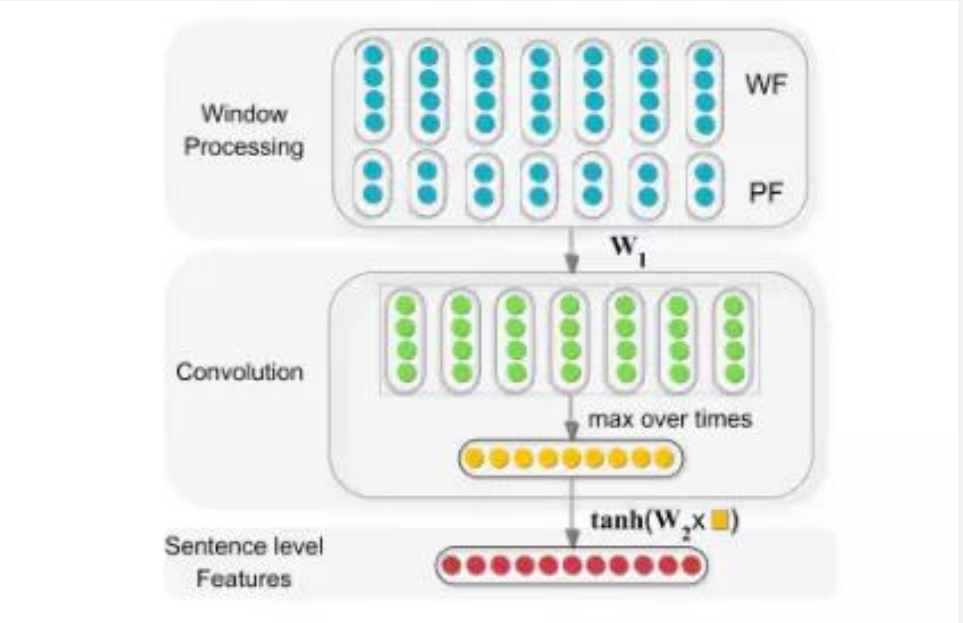
$$P(pattern_i) = \frac{C_a}{C_o}$$

其中C_a表示该集合中成功匹配$pattern_i$、并且<ANSWER>对应部分正好是答案实体的句子数目，C_o表示该集合中成功匹配$pattern_i$的句子数目。

4.3、基于监督学习的方法——CNN/RNN







Zeng, . (2014). Relation Classification via Convolutional Deep Neural Network. Coling, 2335 - 2344

4.3、基于监督学习的方法——CNN/RNN

Lexical—Feature

L1：entity1

L2：entity2

L3：entity1的左右两个tokens

L4：entity2的左右两个tokens

L5：WordNet中两个entity的上位词

	Feature Sets	F1
Lexical	L1	34.7
	+L2	53.1
	+L3	59.4
	+L4	65.9
	+L5	73.3
Sentence	WF	69.7
	+PF	78.9
Combination	all	82.7

Zeng, . (2014). Relation Classification via Convolutional Deep Neural Network. Coling, 2335 - 2344

4.4、半监督学习的方法——自举

算法流程：

1、使用某个关系r对应的有限标注数据（即满足该关系的实体对集合），对无标注文本进行实体标注。

2、从标注结果中抽取出r对应的关系模板

3、将新抽取出的模板应用到无标注文本上，获取更多满足关系r的实体对

4、重复上述过程，直到到达预先制定的停止条件

DIPRE：dual interative pattern relation expansion

Sergey Brin. Extracting Patterns and Relations from the World Wide Web[M]. Berlin:springer, 1999.

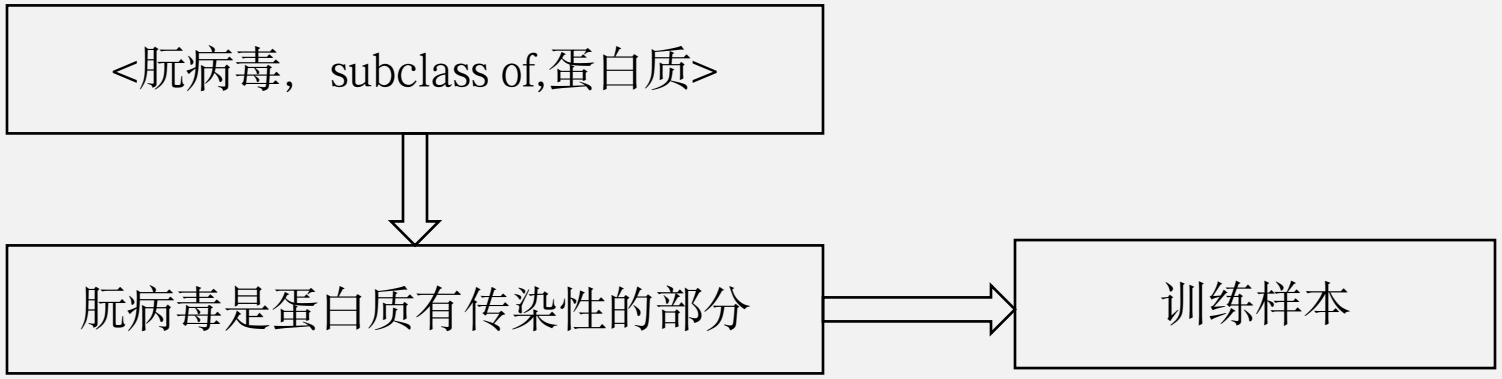
Snowball：

Eugene Agichtein, Luis Gravano. Snowball：Extracting Relations from Large Plain—Text Collections[C]. Acm Conference on Digital Libraries, 2000

4.4、半监督学习的方法——远程监督

- **主要思想**：将已有的**知识**对应到丰富的非结构化语料中从而生成大量的训练数据

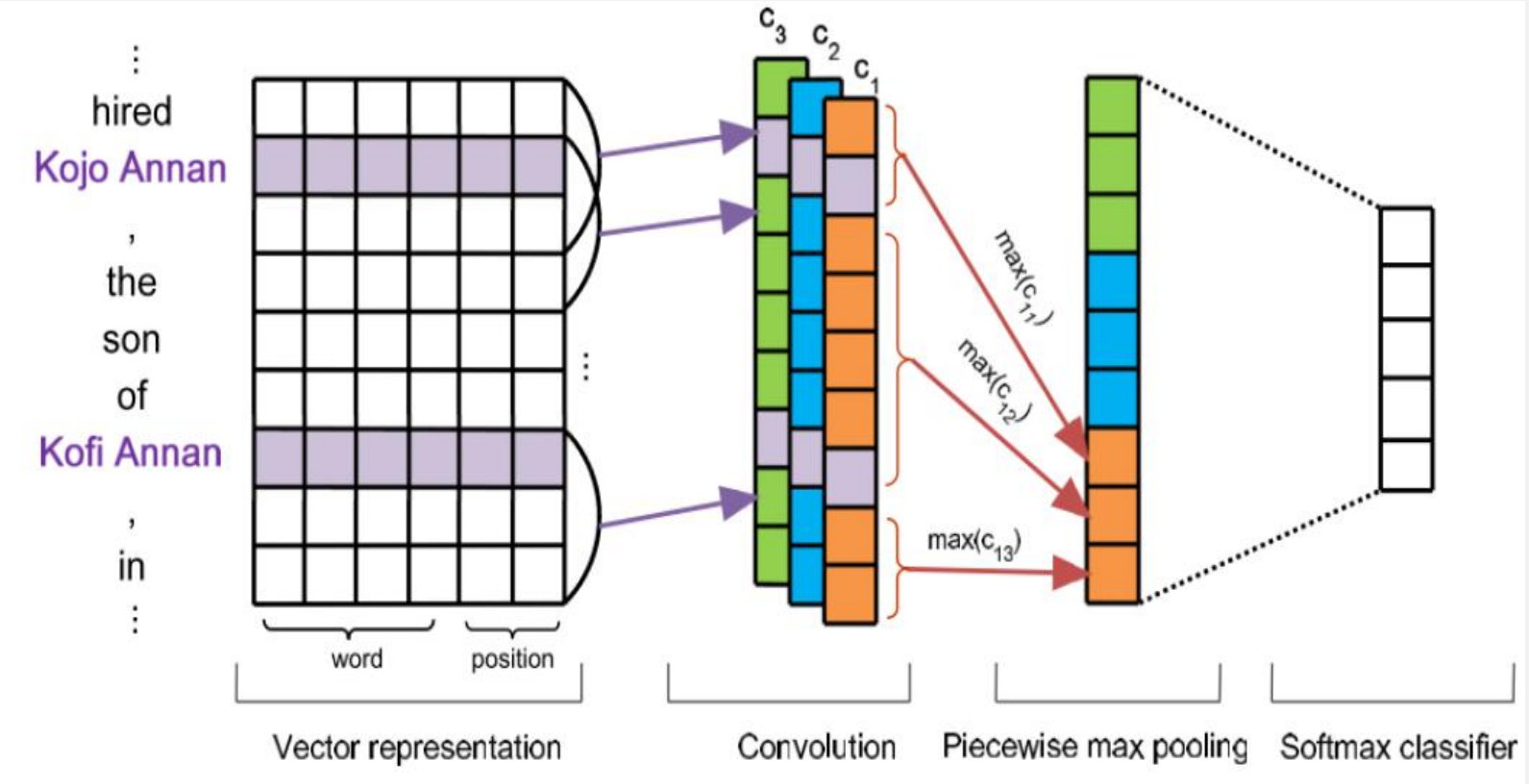
- **知识来源**：人工标注、现有的知识库、特定的语句结构



Distant supervised 会产生有大量噪音或者被错误标注的数据，直接使用supervised的方法进行关系分类，效果很差。

4.5、基于监督学习的方法——PCNN

Zeng (2015). Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. EMNLP



4.6、代码环节

准备动手

https://github.com/buppt/ChineseNRE


谢 谢 大 家
Multipurpose Presentation