远程监督人物关系抽取中的去噪研究

黄蓓静 贺 樑 杨 静
(华东师范大学计算机科学技术系 上海 200241)

摘 要 在远程监督人物关系抽取算法中,训练集的构造是利用知识库对齐自由文本的方式完成的。这一过程会产生部分共现句无法表达当前实体对关系的情况。分析此问题产生的原因,结合词法、句法两个方面,提出一种利用句子模式聚类及模式评分对远程监督人物关系抽取过程训练集进行去噪的方法。该方法首先利用词向量生成特定关系描述候选词,其次针对关系描述候选词提取句子模式并进行模式聚类,最后对模式聚类结果进行评分。通过筛选评分较低模式对应句子去掉对关系描述能力不强甚至无法描述关系的句子,得到过滤后的训练集。实验证明,利用该方法对不同语料库进行远程监督原始训练集句子过滤能够在保证召回率的情况下取得 3% ~5% 准确率提升。

关键词 人物关系抽取 远程监督 模式提取 模式聚类 模式评分

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2017.07.003

RESEARCH ON NOISE REDUCTION IN DISTANT SUPERVISED PERSONAL RELATION EXTRACTION

Huang Beijing He Liang Yang Jing
(*Department of Computer Science and Technology, East China Normal University, Shanghai 200241, China*)

Abstract Distant supervised personal relation extraction heuristically aligns knowledge base with free text to construct training data set. This process can result in invalid aligned sentences which can't express relation of the given entity pair. In this paper, we analyse the reasons of this problem, combined with the lexical and syntactic aspects, and propose a method to denoise the training data set of distant supervised personal relation extraction by using sentence pattern clustering and pattern scoring. Firstly, we generate relation candidate feature words with the help of word embedding. Then, we adopt an algorithm for sentence pattern extraction and pattern clustering by using relation candidate feature words. Finally, we compute confidence of pattern clusters and discard sentences which pattern confidence below threshold. By filtering low sentences with pattern confidence, we can discard sentences with low or no ability to describe specific relation and get the filtered training data set. Experiments show that this method can be used to monitor the different corpus remotely. The original training data set can achieve 3% to 5% accuracy improvement under the condition of guaranteeing the recall rate.

Keywords Personal relation extraction Distant supervised Pattern extraction Pattern clustering Pattern scoring

0 引 言

关系抽取是自然语言处理研究的热点之一,人物关系抽取是关系抽取的一个重要研究方向。

关系抽取方法主要分为有监督、无监督、半监督三类。为了解决有监督方法构造训练集耗费人力及无监督方法准确率不高的缺点,Mintz 等人于 2009 年提出的远程监督关系抽取方法[1]。

Mintz 等人利用 Freebase 作为知识库,将 Freebase 丰富的关系对齐到 Wikipedia 文本中,获取同时出现关系对实体句子(下文称共现句),构成训练集正样本,

收稿日期:2016 - 08 - 15。上海市经济和信息化委员会项目(150643);闵行区人才发展专项资金。**黄蓓静**,硕士生,主研领域:数据挖掘,关系抽取。**贺樑**,教授。**杨静**,副教授。

并随机抽取 Freebase 中不存在关系的实体对对齐得到共现句作为训练集负样本,构成训练集。这既解决了有监督方法训练样本标注大量花费人力的问题,又一定程度上避免了无监督方法准确率较低的问题。

然而,远程监督方法在构造训练集中利用了一个条件很强的基本假设:如果两个实体对存在某种关系,那么任何包含这两个实体的句子都表达了它们的这种关系。实际上,这个假设并不一定成立。这里以人物关系抽取任务举例。如图 1 所示,“赵子琪”和“路金波”是知识库中一对关系实体,关系类型是“夫妻”,通过远程监督方法对齐得到该实体对的三个共现句。其中共现句 1 能够表明人物关系,而共现句 2、共现句 3 虽然同时出现了实体“赵子琪”和“路金波”,但语义上却不能表达他们之间的“夫妻”关系。这种包含两个实体但不表明实体间关系的无效共现句会影响训练集的质量,作为噪声数据,应该予以去除。

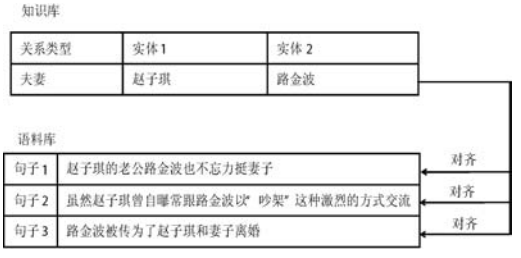

图 1 远程监督人物关系抽取过程中产生无效共现句的一个例子

从图 1 给出的例子可以分析得到噪声产生的原因有:

1) 共现句中不包含能够描述人物对关系的关系描述词,如共现句 2;

2) 虽然共现句包含描述关系的关系描述词,但此关系描述词描述的并非实体 1 和实体 2 之间的关系,如共现句 3 中虽然出现了描述“夫妻”关系的关系描述词“妻子”,但“妻子”并非描述路金波和赵子琪之间的关系。

针对原因 1,我们可以从词法的角度进行解决。近年来,深度学习在 NLP 领域取得了巨大进展,Bengio 等[2]提出的三层神经网络语言模型训练出的词向量在一定程度上包含了词语含义。根据 Mikolov 等[3]研究表明,词向量间存在关系:vec(King) - vec(Man) + vec(Woman) ≈ vec(Queen)。基于词向量我们可以计算出词语之间语义相似度,文献[4]对使用词向量进行同义词扩展进行了研究。关系共现句中关系描述词通常和关系类型本身有密切的语义联系,比如关系类型“夫妻”和关系描述词“结婚”在词向量表示下具有较高相似度。基于此,我们设计了一种基于词向量筛选关系描述候选词的方法。我们首先利用百度百科、互动百科等中文在线百科抓取的关系类型整合出预定义关系描述词集合。然后利用语料训练出词向量。最后筛选出共现句中词频大于阈值且与预定义关系描述词的词向量相似度大于阈值的动词和名词,作为关系描述候选词。

对于原因 2,结合共现句 3 可以看出,对于包含相同关系描述词的不同句子,由于句式结构不同,导致对关系描述强弱程度不同。句子模式抽取方法广泛应用于关系抽取特别是开放式领域关系抽取中[5]。Snowball[6]系统对由种子实例得到的关系共现句提取模式,并对模式评分,使用置信度较高的模式抽取出新的关系实例。文献[7]在对于关系共现句提取模式的基础上进行句子模式相似度、句子模式泛化,实验结果表明,利用泛化后的句子模式进行关系抽取取得了较高准确率。受此启发,我们引入了句子模式提取、句子模式相似度计算、句子模式聚类以及句子模式聚类结果评分等手段,对关系描述较弱或者不能描述关系的句子模式对应的句子进行去除,完成训练集的去噪。

1 相关工作

1.1 远程监督

文献[8]首次在解决生物信息学领域问题时提出了远程监督方法,而后文献[1]首次将远程监督方法引入关系抽取任务中。文献[1]将 Freebase 丰富的关系实例映射到 Wikipedia 文本中,获取了大量的训练样本,从而有效解决了有监督关系抽取方法标注训练语料需要耗费大量人力物力及无监督关系抽取方法准确率较低的缺点。

文献[9]尝试将 90 多种不同关系类型的 DBpedia 关系实例映射到 Wikipedia 文本中,得到 100 000 多个训练样本,使用 logistic 回归训练分类器,关系抽取结果 F 值达 80%。文献[10]首次尝试利用中文在线资源(互动百科)构建中文人物关系抽取系统。该系统利用互动百科抓取到的人物关系对齐多种中文自由文本语料(百度、搜狐、搜狗中文新闻语料),以此获取训练样本,采用标签传播算法训练模型,得到 68% 左右的准确率。这充分说明了远程监督方法在人物关系抽取任务中的适用性和有效性。但以上文献并没有对远程监督方法基本假设引入的噪声训练数据进行去噪处理,某种程度限制了模型准确率的进一步提升。

1.2 关系抽取中的句子模式提取

句子模式抽取是关系抽取中一种重要方法[7]。相

关学者对应用句子模式提取的关系抽取进行了大量的研究。文献[5,11－12]利用人工定义正则表达式方式对上下位关系、整体部分关系进行了抽取。文献[13]提出了一种基于序列模式挖掘的人物关系抽取方法。文献[14]利用模拟退火算法挖掘 Web 中人物关系的最小描述模式集合,利用得到的模式进行关系抽取。文献[7]通过对利用 WordNet 对齐 Wikipedia 得到的共现句替换实体 1、实体 2 为特殊标志及为其他词语加上词性标注得到模式。此后利用最小编辑距离对模式进行泛化,最后利用泛化后模式进行新的关系实例抽取。文献[15]在文献[7]的基础上进行了改进,预先提取关系描述词,并在最小编辑距离计算时代价函数加入词性的考虑。Snowball[6] 也运用了模式提取的方式来生成更多种子实例。Snowball 提出了一种模式置信度计算方式来过滤"不可靠"模式,防止"不可靠"模式提取得到"不可靠"种子实例。但 Snowball 系统的句子模式提取方法过于简单,对于语义相近、结构相似的句子模式没有进行合并,句子模式泛化性不强。

句子模式描述了句式特点,一定程度上体现了句子的语义。本文结合文献[7,15],提出了一种利用句子模式抽取、聚类的方法对远程监督方法中的共现句进行去噪。由于文献[7,15]提出的句子模式提取方法没有对模式进行评分,本文结合文献[6]中对句子模式置信度评分的方法对句子模式聚类结果进行进一步过滤,得到关系表述能力较强的句子模式。

2　远程监督人物关系抽取去噪算法

针对第 0 节分析的噪声共现句产生原因及对策,我们设计了图 2 所示算法框架进行远程监督人物关系抽取的去噪。

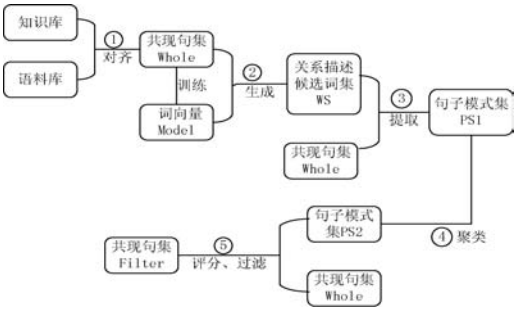

图 2　远程监督人物关系抽取去噪算法框架

如图 2,步骤 1 是远程监督方法的标准步骤,得到的共现句集 *Whole* 是传统远程监督方法不经过去噪得到的训练集。步骤 2 － 步骤 5 组成了本文提出的方法,主要步骤包括关系描述候选词集 *WS* 生成、句子模式提取、句子模式聚类、句子模式聚类结果评分。通过这四个步骤,我们将共现句集 *Whole* 过滤得到共现句集 *Filter* ,即本文方法得到的最终产物:经过筛选得到的高质量训练集。

本节余下内容将对框架上步骤 2 － 步骤 5 这四个关键环节分别展开进行详细阐述,分别是:基于词向量的关系描述候选词生成方法(对应步骤 2)、句子模式提取(对应步骤 3)、句子模式相似度计算及聚类(对应步骤 4)、句子模式置信度计算及句子筛选(对应步骤 5)。后文算法中输入输出符号与本算法框架图采用一致的符号。

2.1　基于词向量的关系描述候选词生成方法

对于特定的关系类型,一般能够找到一系列关系描述词来表达该关系。比如表达"夫妻"关系的关系描述词可能是"丈夫"、"妻子"、"结婚"等;表达"父母/子女"关系的关系描述词可能是"父亲"、"母亲"、"儿子"、"女儿"等。关系描述词一般与关系类型在语义上存在较高相似度,且关系描述词一般在共现句中出现的次数较高。根据此基本假设,我们提出了一种基于词向量的关系描述候选词生成方法。

首先,我们定义词语 w 与关系类型 r 的相似度如下:

$$similarity(w,r) = \frac{\sum_{w' \in RSr} similarity(w,w')}{count(RSr)} \qquad (1)$$

其中 RSr 表示关系类型 r 的预定义关系描述词列表,$similarity(w,w')$ 表示采用词向量表示的词语 w 和预定义关系描述词 w' 的余弦相似度,$count(RSr)$ 表示关系类型 r 的预定义关系描述词个数。

预定义关系描述词可以由知识库的关系类型整合得到,也可以人工定义。由于在线百科的词条是由众多网民共同维护编辑而得,对于同一关系的关系类型描述词并不唯一,这给我们整合预定义关系描述词带来了便利。比如对于"夫妻"关系,不同的描述有"夫妻"、"妻子"、"丈夫"、"老婆"、"老公"等。我们把表述同一关系的不同词语作为该关系的预定义关系描述词,得到表 1 所示预定义关系描述词集 *RS* 。

表 1　由在线百科关系类型整合得到的预定义关系描述词

关系类型	预定义关系描述词
夫妻	妻子、丈夫、b 老公、老婆、夫妻、配偶
父母子女	父亲、母亲、儿子、女儿、母子、父子、母女、父女、大女儿、大儿子
情侣	女朋友、男朋友、女友、男友、情人、情侣
兄弟姐妹	哥哥、姐姐、弟弟、妹妹、胞弟、胞妹、兄长、大哥
同学	同学、同窗
经纪人	经纪人、经理人

例如,我们要计算词语“同班同学”和关系“同学”的相似度,由式(1)及表 1 给出的预定义关系描述候选词,我们可以得到:

$$similarity(w = 同班同学, r = 同学) = \frac{similarity(w = 同班同学, w' = 同学) + similarity(w = 同班同学, w' = 同窗)}{2} = \frac{0.63 + 0.57}{2} = 0.585$$

下面给出基于词向量的关系描述候选词生成算法:

算法 1 基于词向量的关系描述候选词生成算法

输入:共现句集 Whole、预定义关系描述词集 RS、词向量模型 Model、保留词占所有词比例 ratio、相似度阈值 ts、待抽取关系

输出:关系描述候选词集 WS

```
1. FOR EACH 关系 r IN 待抽取关系
2.     FOR EACH 句子 s IN Whole AND s 对齐关系为 r
3.         FOR EACH 词语 w IN s 的中文分词结果列表
4.             IF(w 是动词或者名词)
5.                 记录 w 词频
6.         END FOR
7.     END FOR
8.     按照词频从高到低排序,取前 ratio * 所有词个数个词语组成集合 words
9. END FOR
10. FOR EACH w IN words
11.     IF(Model.similarity(w,r) > ts)
12.         w 加入对应于 r 的关系描述候选词子集 WSr
13. END FOR
14. FOR EACH 关系 r IN 待抽取关系
15.     Wsr 整合至 WS 中
16. END FOR
17. RETURN 关系描述候选词集 WS
```

由于后续算法会进一步剔除某些描述能力弱的关系描述词对应的句子,故算法 1 的主要目的是粗筛选关系描述词,应较全面获取关系描述候选词。由此,算法 1 中的保留词占所有词比例 *ratio* 可适当取 20%,相似度阈值 *ts* 可以适当取最高词语相似度的 0.4 ~ 0.6 倍,以获得较全面的关系描述候选词。

2.2 句子模式提取

对于拥有相同关系描述词的共现句,其表述关系类型的能力不尽相同。比如,共现句 1“冯小刚和徐帆看望了朋友生病的妻子”和共现句 2“作为冯小刚妻子的徐帆”。虽然两个句子都拥有关系描述词“妻子”,但共现句 1 不能表明冯小刚和徐帆的“夫妻”关系,共现句 2 则说明了冯小刚和徐帆是“夫妻”关系。不同句式拥有不同的描述关系能力。我们可以采用句子模式提取的方法来抽象出句式特点。本文提出的句子模式提取方法基于文献[7,15]进行改进。句子模式提取算法描述如下:

算法 2 句子模式提取算法

输入:共现句 sentence、人物实体 1、人物实体 2、共现句对应关系 r、关系候选描述词集 WS

输出:句子模式 pattern

1. 初始化 pattern = sentence
2. 将 pattern 中人物实体 1 和实体 2 分别替换为 < E1 > 、< E2 >
3. 保留 pattern 中对应关系 r 的关系描述候选词子集 WSr 内的词语及词性
4. 保留 pattern 中词“是”、“的”、“与”、“和”、“跟”及其词性
5. 将 pattern 中时间替换成 < TIME >,地点替换成 < LOC >
6. 对于 pattern 中非 2、3、4 步骤中需要保留的词语,只保留词性,词语替换成“.”
7. 只截取保留 pattern 中 < E1 > 、< E2 > 、关系描述候选词之间的句子部分,之外的句子部分丢弃
8. RETURN pattern

根据算法 2,我们可以抽取共现句 1“冯小刚和徐帆看望了朋友生病的妻子”对应的句子模式。句子 1“冯小刚和徐帆看望了朋友生病的妻子”中“妻子”是关系类型“夫妻”描述候选词,经过算法 2 进行句子模式提取可以得到对应模式为“ < E1 > 和/cc < E2 > ./v ./ule ./n ./vi 的/ude1 妻子/n”,共现句 2“作为冯小刚妻子的徐帆”提取出句子模式“ < E2 > 妻子/n 的/u < E1 >”。只保留 < E1 > 、< E2 > 、关系描述候选词之间的句子部分,是出于这么一种考虑:比如共现句“佟大为妻子关悦通过为小两口送上祝福”和共现句“黄磊妻子孙莉也一下成了众人羡慕嫉妒恨的对象”拥有相同的表达夫妻关系的常见模式“A 妻子 B”,如果将句子后半部分的描述也纳入句子模式,这两个本来对夫妻关系具有相同模式的描述各自抽取出的句子模式将具有较大的差别。

2.3 句子模式相似度计算及聚类

利用算法 2 抽取的句子模式已经具有一定的泛化性,比如“佟大为妻子关悦通过为小两口送上祝福”和句子“黄磊妻子孙莉也一下成了众人羡慕嫉妒恨的对象”映射到了相同的句子模式“ < E1 > 妻子/n < E2 > ”。但仍存在相似程度较高的句子模式相对独立的情况,句子模式需要进一步泛化。文献[7]提出使用

最短编辑距离对句子模式进行相似度计算、合并。

最短编辑距离由文献[16]提出,常用于描述字符串之间相似度。最短编辑距离被定义为字符串 A 使用的最少的删除、增加、替换步骤来修改成字符串 B 的步骤次数。最短编辑距离可以使用动态规划来求解,详细求解方法可参考文献[16]。

我们计算出任意两个句子模式之间的最短编辑距离,以此作为两个句子模式的相似度。随后,对于在相同关系类型下拥有相同关系描述候选词的句子模式使用谱聚类的方式进行聚类,使句子模式得到泛化,过程描述如算法 3。

算法 3　句子模式聚类算法

输入:句子模式集 PS1、聚类簇大小 t1、关系描述候选词集 WS、待抽取关系

输出:聚类后句子模式集 PS2

1. FOR EACH 关系 r IN 待抽取关系
2. FOR EACH 关系描述候选词 f IN r 对应关系描述候选词子集 WSr
3. 得到 r、f 对应的句子模式子集

PSrf = {p ∈ PS1 ∧ p 描述关系类型为 r ∧ p 包含关系描述候选词 f}

4. FOR EACH p1 IN PSrf
5. FOR EACH p2 IN PSrf
6. IF(p1 ≠ p2)
7. 计算 p1、p2 最短编辑距离 distance,并令 sim(p1,p2) = distance,构造相似度矩阵 Msim
8. 使用谱聚类算法对 PSrf 内的所有模式进行聚类,簇大小为 t1,结果保存到聚类结果 PS2rf 中
9. END FOR
10. END FOR
11. END FOR
12. END FOR
13. FOR EACH 关系 r IN 待抽取关系
14. FOR EACH 关系描述候选词 f IN WSr
PS2rf 整合至 PS2 中
15. END FOR
16. END FOR
17. RETURN PS2

算法 3 中,为了增强聚类后句子模式的泛化性,聚类簇大小不宜过大。具体可根据聚类前句子模式数量,适当取聚类簇大小 $t1$ 为 3 ~ 5。

2.4　句子模式置信度计算及句子筛选

对于经过算法 2 句子模式提取和算法 3 句子模式聚类得到的句子模式集 PS2,需要引入一定的评分机制来评价不同模式聚类结果描述关系类型能力的强弱。为此,我们借鉴了文献[6]中给出的模式评分方法,句子模式聚类结果置信度评分采用式(3)进行计算:

$$Conf(P_{\text{cluster}}) = \frac{P_{\text{cluster}} \cdot positive}{P_{\text{cluster}} \cdot positive + P_{\text{cluster}} \cdot negative} \quad (2)$$

$$Conf_{RlogF}(R_{\text{cluster}}) = Conf(P_{\text{cluster}}) \times \log(P_{\text{cluster}} \cdot positive) \quad (3)$$

其中,式(2)采用句子模式聚类结果 P_{cluster} 对应的句子正例占所有句子的比例作为 P_{cluster} 的置信度。式(3)对式(2)进行了扩展,在置信度计算中加入了句子模式常见程度,认为常见的句子模式具有更高的置信度。

我们在语料中寻找符合句子模式 P_{cluster} 的共现句,如果共现句实体对关系和知识库一致,则判为正例;若共现句实体对关系和知识库矛盾,则判断为负例。

比如某个表示“父母子女”关系的句子模式簇里面拥有模式“ < E1 > ./v 母亲/n < E2 > ”,对应共现句“房祖名曝成龙不准母亲林凤娇穿尖头鞋及露趾鞋”。从知识库得知,“成龙”和“林凤娇”是“夫妻”关系,不是“父母子女”关系,所以计入该簇的一个负例。而共现句“侯佩岑携母亲林月云亮相上海”也符合上述句子模式,由知识库知“侯佩岑”和“林月云”是“父母子女”关系,此时该簇的正例数加 1。通过统计每个句子模式簇的正负例数量,可以计算出该句子模式簇的置信度评分。借助评分,我们设计了算法 4 筛选出“质量好”的共现句留下作为训练集。

算法 4　句子筛选算法

输入:共现句集 Whole、保留句子比例 ratio、句子模式簇评分结果 ConfPattern、待抽取关系输出:筛选后的共现句集 Filter

1. FOR EACH r IN 待抽取关系
2. 在共现句集 Whole 中获取关系 r 对应的共现句子集 Wholer
3. 保留句子数量 countSentence = Wholer 句子数量 count(Wholer) * ratio
4. 对 Wholer 内句子抽取对应 pattern,以 pattern 的置信度评分 conf 作为句子对应的置信度评分 conf
5. 按照句子置信度评分 conf 从大到小对 Wholer 内句子进行排序,抽取前 countSentence 个句子放入关系 r 过滤后的共现句子集 Filterr
6. END FOR
7. FOR EACH r IN 待抽取关系
8. Filterr 整合至 Filter 中
9. END FOR
10. RETURN Filter

借助算法 4,我们可以把对齐后得到的共现句集 Whole 经过筛选,得到描述关系类型能力较强的句子,组成共现句集 Filter,作为训练集。

3 实 验

3.1 数据集

作为两大中文在线百科,互动百科[18]及百度百科[19]由数量众多的用户参与编辑,词条资源丰富。我们首先由新浪娱乐明星资料页[20]获取部分明星姓名作为人物种子,而后在百度百科、在线百度获取对应人物关系,并把不在种子列表的明星姓名加入种子列表,进行迭代,扩充我们的知识库。最终获取的知识库共有人物 7 954 人,人物对 19 472 对,关系描述词 1 078种。

我们采取爬虫抓取的方式获取语料库。我们抓取了新浪娱乐新闻网、搜狐娱乐新闻网、腾讯娱乐新闻网等大型新闻网近期娱乐新闻,文本文件共计 400 MB。

我们挑选出 6 种常见关系(夫妻、父母子女、兄弟姐妹、情侣、同学、经纪人)进行实验。知识库中 6 种关系的人物对共有 3 963 人,我们按照 6∶4 的比例划分出训练集和测试集。训练集对齐得到共现句集 *Whole* 情况见表 2 所示。

表 2 训练集 Whole 人物对数量、句子数量

关系类型	人物对数量	共现句数量
夫妻	669	73 262
父母子女	650	11 891
兄弟姐妹	713	2 587
情侣	318	16 351
同学	323	2 067
经纪人	38	846

3.2 实验及评价

3.2.1 训练集设置、模型及特征选择

实验目的在于对比本文方法筛选出的句子构成的训练集是否对训练模型性能有所提升。我们设计了两个对比训练集,一个是采用原始共现句集 *Whole*,并随机挑选不存在关系的实体对对齐得到共现句作为关系类型“其他”,加入共现句集 *Whole* 形成训练集 *Whole*。另一个采用筛选后的共现句集 *Filter*,同样加入与训练集 *Whole* 相同的关系类型为“其他”的共现句。由表 2 可知,不同关系类型的共现句数量差别较大,为了使样本比例不过于悬殊,我们对训练集 *Whole* 进行了采样。对共现句数量大于 5 000 的关系类型,随机采样5 000 个共现句,数量小于等于 5 000 的关系类型,保留所有共现句,组成最终的训练集 *Whole*。另外,我们由共现句集 *Filter* 和训练集 *Whole* 求交集得到训练集 *Filter*。训练集 *Filter* 相当于在训练集 *Whole* 的基础上进行句子筛选得到。

本文采用 HanLP 中文分词工具[21]进行中文分词,并使用 gensim[22]进行词向量训练。

特征采用关系抽取任务常用特征[17],即实体 1、实体 2 前后 w 个窗口大小词语、词性。根据文献[17],w 取 2 效果最好,这里,我们取 $w = 2$。

对于训练集 *Whole* 和训练集 *Filter* 我们分别使用 libsvm[23]进行模型训练。

为了对比算法 3 里保留句子比例 *ratio* 的不同取值对训练集 *Filter* 的性能影响。我们设置了 3.2.2 节的对比实验,旨在找到使训练集 *Filter* 性能最佳的 *ratio*,分析其原因,并对训练集 *Whole* 和最佳情况下训练集 *Filter* 的性能进行对比。此外,为了验证本文去噪方法的移植性,3.2.3 节对原始训练集 *Whole* 和训练集 *Filter* 在不同语料库上进行对比实验。3.2.4 节给出了训练集 *Whole* 和训练集 *Filter* 训练时间、训练集存储空间、模型存储空间等对比。

3.2.2 不同保留句子比例对筛选训练集性能的影响对比

为了研究算法 3 中不同保留句子比例 *ratio* 对训练集 *Filter* 性能影响,我们对算法 3 中不同保留句子比例 *ratio* 分别尝试了 7 组不同取值:0.2、0.4、0.5、0.7、0.8、0.9、0.95。以横坐标代表 *ratio* 的不同取值,纵坐标代表对应的准确率、召回率、F 值,水平虚线代表训练集 *Whole* 该指标的数值,用于比较。绘制出 *ratio* 在不同取值下关系抽取结果的准确率、召回率、F 值情况,得到图 3 – 图 5。

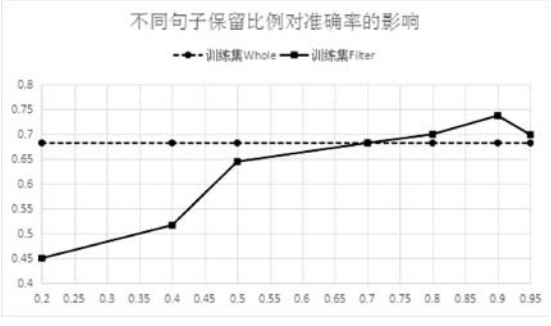

图 3 不同句子保留比例下的准确率

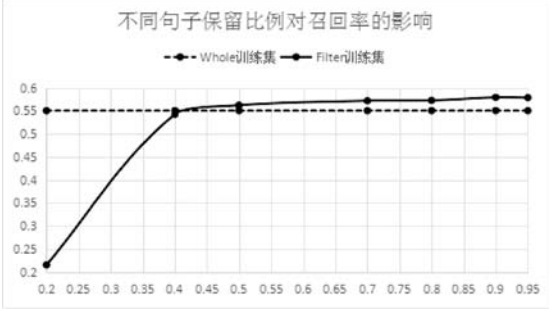

图 4 不同句子保留比例下的召回率

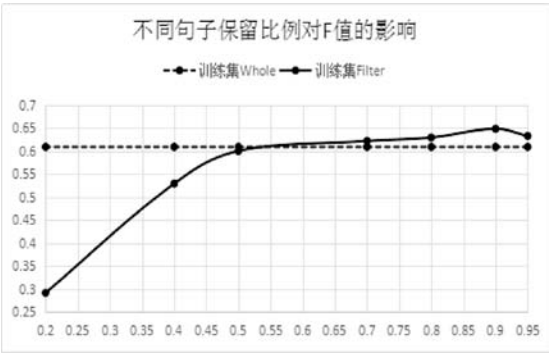

图 5　不同句子保留比例下的 F 值

由图 3 可知,当 *ratio* 小于等于 0.9 时,随着句子保留比例 *ratio* 的增大,准确率逐渐增加,到达 0.9 时最大,往后随着 *ratio* 增加准确率不再升高,反而略有下降。另外,当 *ratio* 小于 0.7 时训练集 *Filter* 的准确率不如训练集 *Whole*。这是因为保留的句子过少,训练样本不足以涵盖足够多的关系表达模式,且样本数量不足,导致训练得到的模型容易过拟合,从而准确率不高。当 *ratio* 大于等于 0.7 时,去除部分噪声共现句能得到质量更高的训练集 *Filter*,而从提高训练模型的准确率。但当 *ratio* 大于 0.9 时准确率会有所下降,这是因为去除的噪声共现句不足导致的。当 *ratio* 等于 0.9 的时候准确率最高,此时丢弃的句子大部分为噪声共现句,并且保证了训练样本数量,所以准确率最高。

由图 4 分析可知,当句子保留比例 *ratio* 达到 0.4 时,训练集 *Filter* 能够达到和训练集 *Whole* 接近或更高的召回率,并且随着 *ratio* 增大,召回率保持稳定。这从另一个侧面说明用本文方法进行训练集去噪能够在保持召回率不降低的情况下提升准确率。

由图 5 分析得,在句子保留比例 *ratio* 大于等于 0.5 的情况下,训练集 *Filter* 能够获得比训练集 *Whole* 更高的 F 值。在 *ratio* 小于等 0.9 时,随着 *ratio* 的增大,F 值逐渐增加,当 *ratio* 达到 0.9 时 F 值最大,此后 *ratio* 增加不会再提高 F 值。这是因为 *ratio* 大于等于 0.5 以后,召回率基本稳定不变,而准确率获得逐步提升,至 *ratio* = 0.9 时准确率最高。由此可以得到结论,在句子保留比例小于等于 0.9 时 F 值逐步升高,往后 F 值略有降低。*ratio* 等于 0.9 时能获得最佳模型。

由上面不同句子保留比例准确率、召回率、F 值对比,我们得出句子保留比例 *ratio* = 0.9 的时候训练集 *Filter* 性能较佳。我们对比此时训练集 *Filter* 和原始训练集 *Whole*,统计得到表 3 所示的训练集 *Whole* 和训练集 *Filter* 对应模型比较情况。

表 3　训练集 *Whole* 和训练集 *Filter* 对应模型比较(*ratio* = 0.9)

关系类型	Whole 准确率	Filter 准确率	Whole 召回率	Filter 召回率	Whole F 值	Filter F 值
综合	0.681	**0.736**	0.551	**0.580**	0.609	**0.649**
夫妻	0.779	**0.804**	0.614	**0.715**	0.686	**0.757**
父母子女	0.613	**0.686**	0.781	0.781	0.687	**0.730**
兄弟姐妹	0.726	**0.754**	0.592	**0.605**	0.652	**0.672**
情侣	0.708	**0.783**	0.210	**0.222**	0.324	**0.346**
同学	0.651	**0.697**	0.524	0.515	0.581	**0.592**
经纪人	0.704	**0.731**	0.826	0.826	0.760	**0.776**

由表 3 可知,对于所有六种关系训练集 *Filter* 训练模型较训练集 *Whole* 准确率都获得了 3% ~ 8% 的提升。与此同时,只有“同学”关系的召回率略微下降,其他关系召回率上升或者保持不变。六种关系的 F 值均获得上升。这说明,在合理选择句子保留比例的情况下,经过本文方法过滤句子得到的训练集 *Filter* 较原始训练集 *Whole* 得出的模型能够在保证召回率的情况下,提升关系抽取的准确率,从而提升关系抽取任务 F 值。

3.2.3　不同语料库准确率、召回率、F 值对比

为了验证本文提出的方法是否具有移植性,我们对不同语料库采用本文提出方法进行对比实验。我们对比了爬虫抓取的新闻、维基百科中文语料两种不同中文语料。对于不同语料库我们采用相同的预处理方法、相同的阈值进行实验,汇总六种关系的准确率、召回率、F 值,得到表 4 结果。

表 4　不同语料库下训练集 Whole 和训练集 Filter 对应模型比较

语料库名称	爬虫抓取新闻	维基百科中文语料
语料库对齐后测试集句子数量	164 781	24 389
Whole 准确率	0.681	0.598
Filter 准确率	**0.737**	**0.630**
Whole 召回率	0.551	0.272
Filter 召回率	**0.580**	**0.283**
Whole F 值	0.609	0.374
Filter F 值	**0.649**	**0.391**

由表 4 分析得出,对于不同的语料库,使用本文提出方法进行训练集句子筛选都能达到在保证召回率的情况下提升准确率的目标,从而提升 F 值。同时可以看到,由于维基百科中文语料对齐知识库得到的测试集句子数比爬虫抓取新闻语料少很多。所以在准确率

及召回率上维基百科中文语料较爬虫抓取新闻语料有所下降,其中召回率下降较为严重。这给我们启发:互联网上有大量冗余自由文本,我们可以充分利用互联网丰富的文本资源,扩大语料库规模,而从提高召回率。

3.2.4 同语料库准确率、召回率、F 值对比

使用本文提出方法进行去噪,不仅能提升模型的准确率,还能减少样本量、节约训练集存储空间及训练模型存储空间。以爬虫抓取新闻语料为例,我们统计了训练集 *Whole* 及训练集 *Filter* 在训练时间、存储空间等项目的对比情况,汇总得到表 5。

表 5 训练集 *Whole* 和 *Filter* 存储空间、训练时间等比较

比较项目	训练集 Whole	训练集 Filter	缩减百分比
训练集样本总条数	25 500	15 240	40.24 %
训练集存储空间大小	1 530 KB	984 KB	35.69 %
训练时间	2 361 秒	346 秒	85.35 %
模型储存空间大小	1 810 KB	897 KB	50.44 %

由表 5 可以看出,训练集 *Filter* 较训练集 *Whole* 样本总数减少,从而节约了模型训练时间、模型存储空间。

4 结 语

本文提出了一种利用词向量及句子模式抽取、聚类及评分的方法对远程监督人物关系抽取过程语料库对齐知识库得到的原始训练集中噪声句子过滤,去掉对特定关系描述能力弱甚至不能描述关系的句子,达到对远程监督人物关系抽取方法训练集去噪目的。我们首先通过使用词频及词向量相似度信息确定了待抽取关系的关系描述候选词;接着我们借助关系描述候选词对句子进行了模式抽取,利用最短编辑距离进行句子模式相似度计算,利用谱聚类方式进行句子模式聚类;最后引入置信度评分对句子模式聚类结果进行置信度评分,保留置信度评分较高的句子作为训练集。实验证明,本文提出的远程监督人物关系抽取去噪方法能对原始训练集中关系描述较弱的句子进行提出,达到保证召回率情况下提高准确率的目标,缩减训练集大小和训练时间,提升了模型性能。

但远程监督在知识库和语料库对齐获取共现句的过程,如何合理地分割句子是一个难点。句子分割粒度太细(如用逗号分割),会出现共现句数量减少的情况,但句子分割粒度太粗,则会出现共现句中实体对距离过远,描述关系的句子模式不显著的情况,下一步工作我们将对此展开研究。

参 考 文 献

[1] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2. Association for Computational Linguistics, 2009: 1003 - 1011.

[2] Bengio Y, Schwenk H, Senécal J S, et al. Neural probabilistic language models[M]. Innovations in Machine Learning. Springer Berlin Heidelberg, 2006: 137 - 186.

[3] Mikolov T, Yih W, Zweig G. Linguistic Regularities in Continuous Space Word Representations[C]//HLT-NAACL, 2013: 746 - 751.

[4] 张为泰. 基于词向量模型特征空间优化的同义词扩展研究与应用[D]. 北京邮电大学, 2015.

[5] Hearst M A. Automatic acquisition of hyponyms from large text corpora[C]//Proceedings of the 14th Conference on Computational Linguistics-Volume 2. Association for Computational Linguistics, 1992: 539 - 545.

[6] Agichtein E, Gravano L. Snowball: extracting relations from large plain-text collections[C]//ACM Conference on Digital Libraries. ACM, 2000: 85 - 94.

[7] Ruiz-Casado M, Alfonseca E, Castells P. Automatic extraction of semantic relationships for wordnet 15[M]. Natural Language Processing and Information Systems. Springer Berlin Heidelberg, 2005: 67 - 79.

[8] Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources[C]// PubMed, 1999: 77 - 86.

[9] Nunes T, Schwabe D. Building Distant Supervised Relation Extractors[C]//IEEE International Conference on Semantic Computing. IEEE Computer Society, 2014: 44 - 51.

[10] 潘云, 布勒布丽汗·伊沙巴依, 杨静, 等. 利用中文在线资源的远程监督人物关系抽取[J]. 小型微型计算机系统, 2015, 36(4): 701 - 706.

[11] Hearst M A. Automated discovery of WordNet relations[M]// WordNet: an electronic lexical database. MIT Press, 1998: 131 - 153.

[12] Berland M, Charniak E. Finding parts in very large corpora[C]//Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, 1999: 57 - 64.

[13] 李丹, 罗智勇. 基于序列模式挖掘的人物关系识别[C]// 中国计算机语言学研究前沿进展(2007 - 2009), 2009: 582 - 587.

(下转第 31 页)

发布,产程数据的展示等功能,如图 5、图 6 所示。

图 5　产程数据录入与审核发布

图 6　产程信息的展示

产程数据录入功能采用 DWR 调用医院信息系统中的病人基础数据。DWR 是一个 JavaEE 的 Ajax 框架,允许 Javascript 代码调用远程 Java 的方法[10],并通过 D-Tree 进行产妇信息树状结构实时展示。助产师可以很方便地通过产妇姓名关联到产妇的基本信息并进行产程信息的填写,填写完产程信息后进行保存操作。在保存前,审核与发布状态的按钮是无法点击的,保存后,责任人将对产程信息进行审核与发布。

产程信息的实时展示是以网页的形式,方便在不同介质(手机或电视)展示给产妇和家属。网页采用分栏显示,左边显示产房提供的视频资料,右边提供产妇产程信息的红色字体向上滚动播放;右边网页会在每 5 分钟同步一次产妇的产程信息,让产妇和家属及时获取最新的产程信息。

3　结　语

产程实时汇报系统从 2013 年初在我院投入运行至今,系统满足了产程汇报的准确性、及时性和系统维护方便性的要求。提高了工作效率与护理质量,收到了产妇及家属满意的效果。自 2013 年以来,新生儿出生人数达 3 万多,随着新生儿出生数量的增多,数据的查询性能受到了影响,我们通过增加索引列和数据查询语句优化,有效地提高了数据的查询速度。此外 B/S 架构环境下,需要考虑多浏览器的兼容性。下一步的工作,还将考虑数据安全性问题,以及在数据加密及数据加密后带来的数据查询和数据更新性能优化权衡等方面做进一步地研究。

参　考　文　献

[1] 陈小玲. 产妇产程汇报表在产时的应用及效应探讨[J]. 中国医院统计,2016, 20(3):202 - 205.

[2] 李佳. 产程观察中应用孕产妇候产通报表的效果观察[J]. 临床医学,2015, 23(1):44 - 45.

[3] 康秀兰. 孕产妇候产通报表在产程观察中的应用及效果分析[J]. 国际护理学杂志,2012, 31 (1):70 - 72.

[4] 高雅娜,高伟,窦红哲. 护理干预对正常分娩产妇产程进展的影响[J]. 河北医药,2012, 34 (12):1892 - 1893.

[5] 张磊. 基于新生儿疾病筛查信息系统设计与实现[J]. 微型电脑应用,2016,32(5):61 - 63.

[6] 郭庆燕,张敏,杨贤栋. JQuery Ajax 异步处理 JSON 数据实现气象图片的显示[J]. 计算机应用与软件,2016,33(6):20 - 22.

[7] Bruce Eckel. Java 编程思想[M]. 陈昊鹏,译. 4 版. 机械工业出版社,2007.

[8] 张友生. 系统分析师教程[M]. 清华大学出版社,2010.

[9] 李刚. 轻量级 Java_EE 企业应用实战[M]. 3 版. 电子工业出版社,2011.

[10] 李刚. 疯狂 Ajax 讲义[M]. 电子工业出版社,2009.

(上接第 18 页)

[14] 姚从磊,邸楠. 一种基于 Web 的大规模人物社会关系提取方法[J]. 模式识别与人工智能,2007,20(6):740 - 744.

[15] 张苇如,孙乐,韩先培. 基于维基百科和模式聚类的实体关系抽取方法[J]. 中文信息学报,2012,26(2):75 - 81.

[16] Wagner R A, Fischer M J. The String-to-String Correction Problem[J]. Journal of the Acm,1974,21(1):168 - 173.

[17] 车万翔,刘挺,李生. 实体关系自动抽取[J]. 中文信息学报,2005,19(2):1 - 6.

[18] Hudong encyclopedia[EB/OL]. http://www. baike. com/.

[19] Baidu encyclopedia[EB/OL]. http://baike. baidu. com/.

[20] Sina star[EB/OL]. http://ku. ent. sina. com. cn/star/search.

[21] HanLP[EB/OL]. https://github. com/hankcs/HanLP.

[22] gensim[EB/OL]. http://radimrehurek. com/gensim/.

[23] libsvm[EB/OL]. http://www. csie. ntu. edu. tw/ ~ cjlin/libsvm.