软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(9):2088−2101 [doi: 10.13328/j.cnki.jos.004638]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

关系抽取中基于本体的远监督样本扩充*

欧阳丹彤[1,2], 瞿剑峰[1], 叶育鑫[1,2,3]

[1](吉林大学 计算机科学与技术学院,吉林 长春 130012)

[2](符号计算与知识工程教育部重点实验室(吉林大学),吉林 长春 130012)

[3](吉林大学 国家地球物理探测仪器工程技术研究中心,吉林 长春 130026)

通讯作者: 叶育鑫, E-mail: yeyx@jlu.edu.cn

摘 要: 远监督学习是适合大数据下关系抽取任务的一种学习算法.它通过对齐知识库中的关系实例和文本集中的自然语句,为学习算法提供大规模样本数据.利用本体进行关系实例的自动扩充,用于解决基于远监督学习的关系抽取任务中部分待抽取关系的实例匮乏问题.该方法首先通过定义关系覆盖率和公理容积率,来寻找与关系抽取任务关联性大的本体;然后,借助本体推理中的实例查询增加待抽取关系下的关系实例;最后,通过对齐新增关系实例和文本集中的自然语句,达到扩充样本的效果.实验结果表明:基于本体的远监督学习样本扩充方法能够有效完成样本匮乏的关系抽取任务,进一步提升远监督学习方法在大数据环境下的关系抽取能力.

关键词: 远监督;关系抽取;本体

中图法分类号: TP18

中文引用格式: 欧阳丹彤,瞿剑峰,叶育鑫.关系抽取中基于本体的远监督样本扩充.软件学报,2014,25(9):2088−2101. http://www.jos.org.cn/1000-9825/4638.htm

英文引用格式: Ouyang DT, Qu JF, Ye YX. Extending training set in distant supervision by ontology for relation extraction. Ruan Jian Xue Bao/Journal of Software, 2014,25(9):2088−2101 (in Chinese). http://www.jos.org.cn/1000-9825/4638.htm

Extending Training Set in Distant Supervision by Ontology for Relation Extraction

OUYANG Dan-Tong[1,2], QU Jian-Feng[1], YE Yu-Xin[1,2,3]

[1](College of Computer Science and Technology, Jilin University, Changchun 130012, China)

[2](Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education (Jilin University), Changchun 130012, China)

[3](National Engineering Research Center of Geophysics Exploration Instruments, Jilin University, Changchun 130026, China)

Corresponding author: YE Yu-Xin, E-mail: yeyx@jlu.edu.cn

Abstract: Distant supervision is a suitable method for relation extraction in big data. It provides a large amount of sample data by aligning relation instances in knowledge base with nature sentences in corpus. In this paper, a new method of distant supervision with expansion of ontology-based sampling is investigated to address the difficulty of extracting relations from sparse training data. First, an ontology which has a deep link with relation extraction is sought through the definition of cover ratio and volume ratio. Second, some relation instances are added by ontology reasoning and examples of queries. Finally, the expansion of training sets is completed by aligning the new relation instances and nature sentences in corpus. The experiment shows that the presented method is capable of extracting some relations whose training sets are weak, a task impossible by the normal distant supervision method.

Key words: distant supervision; relation extraction; ontology

为了能够使得网络上的自然语言信息变成结构化的形式,方便分析,研究者提出了不同的关系抽取方法.关

* 基金项目: 国家自然科学基金(61272208, 61133011, 41172294, 61170092); 吉林省科技发展计划(201201011)
收稿时间: 2014-01-31; 修改时间:2014-05-06; 定稿时间: 2014-06-09

系抽取是指从文本内容中检测实体之间的明确或者不明确的关系,并且把它们分类.从机器学习的样本获取角度看,主要有 3 类方法用来从文本中抽取关系事实,它们分别是全监督学习[1]、半监督学习[2]和无监督学习[3].全监督学习是指通过人工标注初始的样本数据,然后利用标记过的数据训练分类器,最后用训练好的分类器去识别一个新句子中是否有某两个实体存在某种给定的关系.全监督的学习方法主要包括基于特征的方法和核方法.而半监督学习则是指使用一个非常小的数据种子实例或者模式做引导学习,在大量的文本里面抽取一些新模式,然后再用这些模式抽取新的实例,新实例再去抽取更新的模式,周而复始,最终得到数据.它的具体方法包括 DIPRE,Snowball 和 KnowItAll.无监督学习不需要初始数据集,而是从大量的文本中抽取介于两个实体之间的字符串,然后对那些字符串进行聚集和简化,得到关系字符串.

随着大数据时代的到来,关系抽取任务面临的适用领域更加开放和复杂[4,5].面对海量和异构数据,研究者提出了远监督方法[6].该方法通过启发式地对齐待抽取关系与自然语句,为关系抽取任务提供大量样本[7].以图1中抽取关系/location/country/capital 为例,在知识库中有实例(Germany,Berlin),而在文本集中有句子“Berlin is the capital of Germany, ...”.系统就会自动地将它们相匹配,形成一个训练实例:

{capital(Germany,Berlin),Berlin is the capital of Germany, ...}.

关系 contains 也是重复上述同样的过程,形成训练集中的一个关系实例.

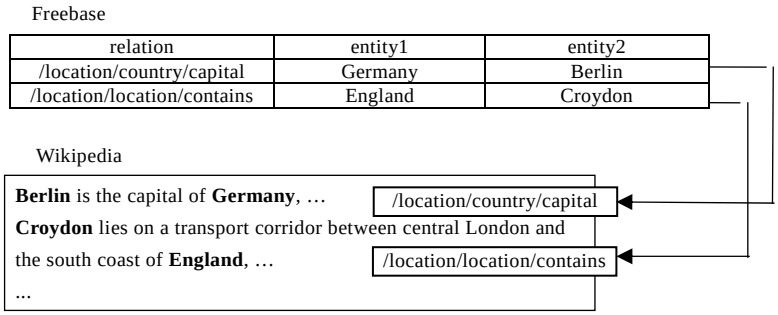

Fig.1 Alignment of relation instance

图 1 关系实例对齐

较以往机器学习方法,远监督学习方法初步解决了大数据的可用性问题[8],使得从以 Wikipedia 为代表的网络大数据中获取实体关系成为可能.如何提升大数据的一致性、精确性、完整性、时效性和实体同一性,使其在远监督学习方法下取得更好的效果,是近年来研究的热点和重点.目前对于 Distant supervision 的研究,一方面是对于初始数据源的不同选取来提升大数据的可用性,比如有人用 Freebase 和 Wikipedia,也有人用 YAGO 和 Wikipedia 进行对应.Mintz 等人[6]提出了将 Freebase 与 Wikipedia 对应来提供样本数据.Freebase 是大型语义数据库,包含有几千种关系类型.对于每个出现在 Freebase 中的实体对,在 Wikipedia 中找到所有包含该实体对的句子集合,来提供文本特征去训练分类器.Mintz 的模型在 102 种关系上抽取了 10 000 个实体,最终的准确率为 67.6%.Nguyen 等人[9]则考虑选取 YAGO 知识库.该方法基于如下的考虑:(1) 关系可以来自不同于 Wikipedia 的语义仓库,例如 YAGO 等;(2) 利用由任何 Wikipedia 文章获取的训练实例.他们证实了训练数据由包含目标关系的命名实体对的句子组成是足够可靠的,并且该方法是目前比较先进的关系抽取模型.运用上面的数据训练分类器,最终 *F*1 达到了 74.29%.这个方法极大地增强了远监督方法在关系抽取方面的应用.另外,实验结果也表明了他们的分类器可以应用到任何一般的文章.

另外一方面的研究侧重于通过数据去噪提升大数据的可用性.因为启发式的数据对齐会在一定程度上产生错误的数据,而且某部分的错误数据可能会对分类器的准确率方面造成严重的负面作用.Shingo[10]提出了减少错误标注的方法,该方法区别于传统的启发式的标注过程.在标注数据的时候,首先会通过自身隐藏的变量去判断赋予的标注是错误的还是正确的.最终的实验结果也表明了这种方法能够有效地检测错误的标签,提高了

样本数据的质量,得到的关系分类器器的效果也会更好.

除了上述研究工作,也有研究者通过运用不同的句子特征来训练分类器[11],例如,有的用词法和句法特征,有的利用动态的辞典特征.对于上述不同的研究方向,都能够在一定程度上提高分类器的准确率.然而仍存在部分实例不完整、不一致等问题,导致分类器的在某些关系抽取上的精度不够令人满意.

本文提出了基于本体的样本扩充方法进行远监督学习下的关系抽取任务.该方法有效地利用了本体的自动推理的功能,将知识库中目标关系与本体进行了映射,然后,通过使用本体的结构,检索出实体对之间的隐含目标关系.在完成启发式匹配后,训练数据得到了扩充.本文将基于本体的样本扩充方法与普通的远监督学习方法及基于多实例的远监督学习方法进行了比较,实验结果表明:在大数据环境下,利用本体推理对数据扩充,进一步提升了样本的一致性、完整性和实体统一性,能有效抽取样本实例缺乏的关系,而这些关系利用普通远监督学习算法是无法抽取的.

1 预备知识

1.1 关系抽取

1.1.1 任务定义

关系抽取任务是检测并且揭示文本中实体之间的语义关系[12].对于每个句子 S,其中包含了实体集合 $\{e_1,e_2,e_3,.....e_n\}$,n 为实体在句子中出现的序号.关系抽取就是识别 e_i,e_j 是否存在某种给定的关系 R_k 记为 $R_k(e_i,e_j)$.例如句子“Cone, a Kansas City native, was originally signed by the Royals and broke into the majors with the team.”我们也许会希望去抽取一种雇佣关系在实体对 Cone 和 Royals 之间.当然,关系抽取也包括多元的关系,即多个实体之间具有某种关系,但是在本文我们不做考虑.

1.1.2 联合特征的选取

为了能够使机器自动地完成这种潜在语义关系的识别任务,我们必须去选取适合的特征描述来表达句子中出现的实体对的知识,从而能够构造出关系的候选.目前,关系抽取所采用的特征有词法、词性、语法、语义等方面的特征.下面比较几种常用的特征[11]:

- 两个实体之间的单词序列(word sequence),例如,句子“Astronomer Edwin Hubble was born in Marshfield, Missouri”中的实体之间单词序列是“was born in”;
- 词性特征(part-of-speech):指标注句子中所有词的词性;
- 两个实体所属的类别(type):PERSON,ORGANIZATION,FACILITY,LOCATION 和 Geo-Political Entity 或者 GPE;
- 依赖解析(dependency parse):根据依赖解析树,找到两个实体之间的依赖路径.对于上面的句子,我们分析得到如图所示的依赖解析,在进一步分析后得到实体之间的依赖路径为(如图 2 所示):

“$\uparrow_{s}$was $\downarrow_{perp}$born $\downarrow_{mod}$in $\downarrow_{pcomp\text{-}n}$”

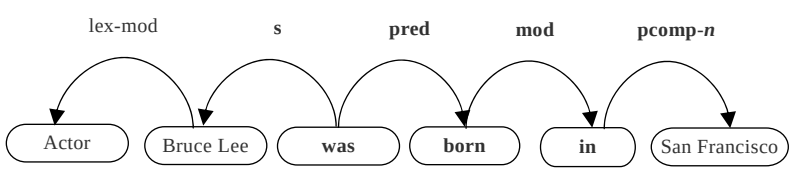

Fig.2 Dependency parse

图 2 依赖解析

本文中,我们并不是单独地采用上述的某种特征,而是利用联合特征,就是将所有的特征同时进行考虑.每个特征都是由句子中的多个属性联合而成,加上命名实体的标记.当两个联合特征相匹配的时候,必须其中的每个特征要素都能精度的匹配.这样的联合特征避免了片面性,提高了准确率.

1.2 远监督学习

远监督学习方法结合了其他几种机器学习方法的优点.Snow 等人[13]在 2004 年使用范式,利用 WordNet 抽取实体之间的上下位(is-a)关系.远监督学习是上述范式的延伸,并且与 Craven 等人使用弱标记生物信息学[14]的数据方法相类似.它使用一个已知的知识库和文本集来提供训练数据,而这种训练数据的产生是基于远监督的内部机制,即任何包含已知知识库中关系的实体对的句子,都是在以某种方式潜在地表达了这种关系.在此条件下,文本集中存在大量的包含给定的实体对的句子,从而可以抽取到大规模特征数据(当然也包括噪声数据).

因此,当全监督训练模式只能使用包含 17 000 个关系实例的小标注语料库作为训练数据的时候,远监督方法已经可以使用更大的数据量——更多的文本、更多的关系以及更多的实例.以 Mintz 的方法为例(知识库:Freebase,文本集:Wikipedia),他的方法使用了 120 多万的 Wikipedia 的文章,102 种关系类型连接了 94 000 个实体,产生了 180 多万的关系实例.另外,分类器结合使用了大规模的特征,有效地排除了噪声特征的影响.远监督方法的数据是受到知识库的监督,而不是被标注的文本,所以它不会遭受到过度拟合和领域依赖的困扰.而且,监督的知识库意味着:与无监督的方法不同,远监督分类器输出的都是使用规范名称的关系.

1.3 本 体

本文使用 OWL(ontology Web language)语言来描述本体,W3C 组织定义了标准的本体语言为 OWL,而 OWL 的逻辑基础是描述逻辑(description logic,简称 DL)[15].描述逻辑(DL)提供了对感兴趣的领域内的实体之间的关系进行建模的方法.描述逻辑有如下 3 种类型的实体:

- 概念(concept):概念表示个体的集合;
- 角色(role):角色表示个体间的二元关系;
- 个体名称(individual):个体名称表示领域内的单个个体.

1.3.1 ABox 公理的断言事实.

ABox 公理获取关于命名个体的知识,就是表达命名个体的所属以及它们之间的如何相互联系.最普遍的 ABox 公理是概念的断言(concept assertions),例如 Father(Blue),表示 Blue 是一位父亲,或者更精确地表示为名字是 Blue 的个体是 Father 概念的一个实例(instance).角色断言(role assertions)描述的是命名个体之间的关系,断言 sonOf(Blue,Jack)表示 Jack 是 Blue 的儿子,或者更精确地表示名字是 Blue 的个体与名字是 Jack 的个体具有 sonOf 上的关系.

1.3.2 TBox 公理术语的知识表达.

TBox 公理是用来描述概念之间的关系,比如说有这样的一个事实,所有的 Father 概念都属于 Parent 这个概念,这种事实可以用概念包含(concept inclusion)的方式来表示 Father⊑Parent,我们称概念 Parent 包含概念 Father,这些知识可以用来推导出个体之间的一些事实.概念等价(concept equivalence)表示两个概念拥有一样的实体,例如 Human≡Person,显然,同义词是等价概念的一种.

例如,构建一个关于家庭成员和他们家族关系领域内的本体,可以使用 Parent 这样的概念来表示所有的父母集合,使用 Female 概念来表示所有的女性成员集合,使用 parentOf 这样的角色来表示父母和他们孩子之间的(二元)关系,使用 julia 和 john 这样的个体名称表示家庭成员 Julia 和 John.而本体中的概念、角色、个体名称分别对应知识库中的类、关系、实体,所以本文利用这种对应的关系进行映射.知识库中的关系及其实例相当于本体中的 ABox 公理断言事实,但是知识库中没有本体中的 TBox 公理,所以不能实现自动推理,推导出隐含的知识.所以对于样本知识库的本体映射,关键的内部机制是系统能够自动地去寻找一个与知识库中的目标关系(即样本中的关系)相映射的本体,这也是本文需要解决的关键问题.

2 基于本体的样本自动扩充

2.1 本体发现

基于本体的样本自动扩充的首要任务就是要找到一个与样本知识库相匹配的本体.该过程类似于本体匹

配、本体映射和基于本体的语义搜索[16],但又与它们有所区别.本体匹配和本体映射是本体到本体的关联,而本文中是知识库到本体公理集的关联.从本体角度看,知识库相当于本体中的断言集 ABox,而本体发现的过程相当于为存在的一个断言集寻找合适的公理集 TBox 的过程.基于本体的语义搜索是通过已知的领域本体搜索相关信息的过程,而本体发现则是与其相反的一个过程,目的在于利用已有信息去获取一个恰当的本体.本体发现的目的是为当前知识库找到一个合适的公理集合组成本体,进而进行有效的本体推理.

我们在借鉴本体匹配、本体映射和基于本体的语义搜索中部分方法的基础上,首先利用语义相似度定义了关系的覆盖度,目的在于期望寻找的公理集合在语义上最接近当前知识库;然后,利用每个角色平均拥有的公理数来定义公理集合的规模,目的在于期望寻找到的公理集合,能够最大限度地利用本体推理演绎出更多关系实例;最后,将关系覆盖度和公理容积率两方面因素综合考虑,给出本体发现的评价标准公式,用于定位适合当前知识库的本体.

2.1.1 关系的覆盖度(cover ratio)

基于本体的样本扩充方法使用相似性比较中的语义相似度作为标准.语义相似度采用分类学的角度(例如 WordNet)比较两种名称的相似度.语义相似度是指名称本身的语义之间的相似程度,而不是简单地去比较两个词汇字母构成的差异.当提高“rivers 和 ditches 是一个意思的时候”,我们不是去比较 rivers 和 ditches 的字符的集合,而是去比较 river 和 ditch 的类属.

Lin[17]在 1998 年提出了语义相似度的比较方法,假设分类以树的形式表示,每个节点表示一个词汇或短语,树中的连线代表连接的两个结点有上下位包含关系.我们用 x_1,x_2 代表需要比较的名称(即树中的某两个节点),C_1,C_2 分别代表 x_1,x_2 的父节点,C_0 则表示的是 x_1 和 x_2 的公共祖先结点.$P(C)$代表附着在 C 的所有子节点占全部结点的比重.则定义 x_1 和 x_2 的相似度的计算公式如下:

$$Sim(x_1,x_2)=\frac{2\times\log P(C_0)}{\log P(C_1)+\log P(C_2)} \tag{1}$$

本文采用基于 WordNet 的 Lin 方法,上述的公式是比较两个名称的相似度.下面我们给出计算两个名称集合的相似度公式:

$$Sim_{set}(S_1,S_2)=\sum_{i=1,y_j\in S_2}^{n}\max(sim(x_i,y_j)) \tag{2}$$

针对于本文,S_1 表示的是样本的类或者关系的集合,而相应 S_2 表示的是本体文件中概念或者属性的集合.最终,我们定义样本被候选本体文件的覆盖度公式如下:

$$Cover(K,O)=w_1\cdot sim_{set}(T,C)+w_2\cdot sim_{set}(P,R) \tag{3}$$

其中,K 表示样本数据,O 表示候选的本体文件,$sim_{set}(T,C)$表示样本中的类被本体中的概念的覆盖程度,$sim_{set}(R,P)$表示样本的关系被本体中的属性的覆盖程度.其中,$w_1+w_2=1$,根据经验分别设定其值为 $w_1=0.2$,$w_2=0.8$.由公式 $cover(K,O)$筛选出候选本体对应的 $cover$ 值最大的本体文件,作为与样本相映射的本体文件来完成进一步的工作.

2.1.2 公理容积率(volume ratio)

显然,在第 2.1.1 节中,样本被本体覆盖得越高,则越可能是一个好的映射.此外,待选取本体公理数量的多少是映射好坏需要参考的另一指标.一般来说,本体公理集规模越大(即所含公理数越多),本体推理时越能体现发现隐含知识的效果.所以,本体公理集中,公理数目是选取本体要考虑的因素之一.但对于本体公理数目占优的本体而言,如果大部分公理描述都集中在其中部分概念和关系上,会在本体推理时导致部分关系的关系实例演绎不足甚至失效.因此,本文通过每个角色平均拥有的公理数来定义公理容积率这个指标,而没有从公理集合的完备性去定义,是因为利用这个指标可以有效定位适于做关系实例扩展的本体.

本文采用有限权图计算公理的容积率,定义有限权图 $G(P,L)$,其中,图中的节点代表本体中的某个属性,图中的边代表属性之间存在公理,边上的权值代表公理的个数.根据候选的本体构造出相应的图,并采用邻接矩阵的方式存储图结构.通过对邻接矩阵的计算,可以得到每个点的平均度数(即每个角色包含的平均公理数),记为

$Vol=(1/|P|)\sum_{l=1}^{|L|}w(l)$.最终,取令 *Vol* 为最大值的本体,认为是公理容积率最高的本体.

2.1.3 联合方法(joint)

该方法综合考虑了关系覆盖度(cover ratio)和公理容积率(volume ratio)这两个方面,避免了单方面考虑的片面性.由此给出计算公式:

$$Mjoint=0.5\cdot Cover(K,O)+0.5\cdot(Vol-\min(Vol))/(\max(Vol)-\min(Vol)) \tag{4}$$

从公式可以看出,该方法同时利用了样本的覆盖度和公理容积率,并且给它们赋予了相同的权重 0.5.在公理容积率计算时,每个角色的平均拥有的公理数相差较大,因此,我们通过找到它们的最大值与最小值,即 max(*Vol*)和 min(*Vol*),利用算式(*Vol*−min(*Vol*))/(max(*Vol*)−min(*Vol*)),使容积率的值落在 0~1 之间.然后,再与覆盖度指标相结合,给出联合方法的评价公式(4).另外,在实际操作中,考虑了关系覆盖和公理容积率两方面因素所发现的本体,排在前几位的语言表达能力基本上都在 SHOIQ 级别,很少存在语言表达能力太弱的本体出现在前面的情况.因为一般来讲,语言表达能力强的本体自然描述领域知识更精细和复杂,其关系覆盖度和概念容积率自然高.因此,可以不用考虑将本体语言表达能力强弱作为发现本体的因素.

2.2 基于本体推理的关系实例扩充

问题定义:本体主要分为概念的推理和 ABox 的推理(ABox 是指获取命名个体的知识,就是表达命名个体的所属以及它们之间的相互联系),而 ABox 的推理问题主要包括一致性检测、实例检测和查询等.基于本体的样本扩充方法利用 ABox 实例查询,即在某个 TBox *T* 的背景下,已知 ABox *A* 和概念 *C*(或角色 *R*),找出所有满足 $A\models R(a,b)$(或 $A\models C(a)$)的实例[18].

内部机制:本体推理工具 Pellet 是基于表算法的推理机.Pellet 会检查知识库的一致性,并且将其他所有的推理服务转化为一致性的检查.基于本体的样本扩充方法是利用 Pellet 解决 SHOIQ 的推理问题,其中主要涉及如下几条规则:

假设(T,R)是一个 *SHOIQ* 的知识库,$\boldsymbol{R}_{T,R}$ 是一组在知识库 T 或者 R 中出现的角色集合以及它们的逆.对于(T,R)的表 *T*,可以定义成三元组($S,\wedge$,E),其中,$\boldsymbol{S}$ 是一组个体,$\wedge:\boldsymbol{S}\to 2^{cl(T)}$映射每个个体到一组概念,而这些概念是 cl(T)的子集,E:$\boldsymbol{R}_{T,R}\to 2^{S*S}$ 映射每一个在 $\boldsymbol{R}_{T,R}$ 中的角色到一组个体对.规则中是 $s,t\in S,C,C_1,C_2\in cl$(T),并且 $R,S\in\boldsymbol{R}_{T,R}$.规则如下:

(P0) if $C\sqsubseteq D\in\mathcal{T}$, then $\mathsf{nnf}(\neg C\sqsubseteq D)\in\wedge(s)$,

(P1) if $C\in\wedge(s)$, then $\neg C\notin\wedge(s)$,

(P2) if $C_1\sqcap C_2\in\wedge(s)$, then $C_1\in\wedge(s)$ and $C_2\in\wedge(s)$,

(P3) if $C_1\sqcup C_2\in\wedge(s)$, then $C_1\in\wedge(s)$ or $C_2\in\wedge(s)$,

(P4) if $\forall R.C\in\wedge(s)$ and $\langle s,t\rangle\in E(R)$, then $C\in\wedge(t)$,

(P5) if $\exists R.C\in\wedge(s)$, then there is some $t\in S$ such that $\langle s,t\rangle\in E(R)$ and $C\in\wedge(t)$,

(P6) if $\forall S.C\in\wedge(s)$, and $\langle s,t\rangle\in E(R)$ for some $R\sqsubseteq S$ with $\mathsf{Trans}(R)$, then $\forall R.C\in\wedge(t)$,

(P7) if $(\geqslant nS.C)\in\wedge(s)$, then $\ulcorner\{t\in S|\langle s,t\rangle\in E(S)$ and $C\in\wedge(t)\}\geqslant n$,

(P8) if $(\leqslant nS.C)\in\wedge(s)$, then $\ulcorner\{t\in S|\langle s,t\rangle\in E(S)$ and $C\in\wedge(t)\}\leqslant n$,

(P9) if $(\leqslant nS.C)\in\wedge(s)$, and $\langle s,t\rangle\in E(S)$, then $\{C,\neg C\}\cap\wedge(t)\neq\varnothing$,

(P10) if $\langle s,t\rangle\in E(R)$ and $R\sqsubseteq S$, then $\langle s,t\rangle\in E(S)$,

(P11) $\langle s,t\rangle\in E(R)$ iff $\langle t,s\rangle\in\mathcal{E}(Inv(R))$,

(P12) if $o\in\wedge(s)\cap\wedge(T)$ for some $o\in C_o$, then $s=t$, and

(P13) for each $o\in C_o$ occurring in *T*, there is some $s\in S$ with $o\in\wedge(s)$.

...

例如,如图 1 的初始样本中有关系实例 capital(Germany,Berlin)和 contains(England,Croydon),在本体结构中,发现 capital 和 contains 这两种角色对应蕴含关系(*capital*⊆*contains*),所以,Pellet 会依据规则(P10)推理出隐含的关系实例 contains(Germany,Berlin),它通过启发式方法匹配的句子集合可作为关系 contains 的训练数据.因此,若 Pellet 推理出的隐含的角色实例 *R*(*a*,*b*),那么我们可以将其作为新增的样本关系实例,用于下一步的启发式匹配.算法 1 实现如下:

算法 1. REASONING.

```
Input: Ontology O;
Output: Relations.
OntModel model=ModelFactory.createOntologyModel(PelletReasonerFactory)   //初始化 Pellet 推理机
model.read(O)                                                  //将本体文件读入推理机中
KnowledgeBase kb=model.getKB( )                                //读取本体中的知识表示的内容
kb.classify( )                                                 //运用推理机,进行知识推理
IndividualIterator Relations=kb.getABox( )                     //获取推理后的知识库中所有的 ABox 断言
return Relations                                               //返回所有的关系及其实例
```

2.3 样本扩充(对齐)

样本扩充的关键任务是扩充训练分类器的特征数据,因此在完成样本知识库中关系实例的增加之后,需要进行启发式匹配过程.启发式匹配是基于远监督学习的前提假设:如果某两个实体具有某种关系,那么任何包含这两个实体的句子就可能表达了这种关系[6].例如,对于关系实例⟨film-director,Steven Spielberg,Saving Private Ryan⟩,在基于假设的条件下,我们可以在 Wikipedia 的自然语言文本中找到如下包含上述的实体对的句子:

1) Steven Spielberg's film Saving Private Ryan is loosely based on the brother's story;
2) Allison co-produced the Academy Award-winning Saving Private Ryan, directed by Steven Spielberg;
3) When Steven Spielberg made Saving Private Ryan he aimed to portray "the terrors and triumphs of D-Day as more than just make-believe";
4) Saving Private Ryan is the most troublesome film in Steven Spielberg's flexography.

下面给出匹配算法 2,算法用伪代码表示.

算法 2. GENERATE_LABELED_DATA.

```
01  DS=∅
02  Sample KB(R): Instances of Relation R
03  for each (Wikipedia article:W)∈Wikipedia
04    S←set of sentences from W
05    for each s∈S
06        ε←set of entities from s
07        for each E1∈ε and E2∈ε and
08            R∈Sample KB
09            if R(E1,E2)∈Sample KB(R)
10          then DS←DS{s,R}
11            end if
12        end for
13    end for
14  end for
15  return (DS)
```

算法 2 中:第 1 行把关系对齐数据初始化;第 2 行~第 14 行遍历每一个文本,并从中找出关系实例所对应的

句子.其中,第 5 行~第 13 行遍历每一个句子,从中找出每个实体对,并判断这个实体对是否存在于知识库的某个关系实例中:如果是,就把该句子加入到该关系实例的对齐数据中去.

3 基于本体的远监督关系抽取任务框架

本文在传统远监督方法的基础上,提出了知识库的本体映射以及本体的自动推理两个步骤.如图 2 所示,虚线部分是本文的主要工作:

(1) 知识库的本体映射,即针对样本中的目标关系,寻求与其相映射的本体文件,使得知识库中的关系、实体结构化;

(2) 本体的自动推理,即利用本体的结构(公理、断言)推导出样本中实体之间隐含的目标关系,从而达到关系实例扩充的效果.

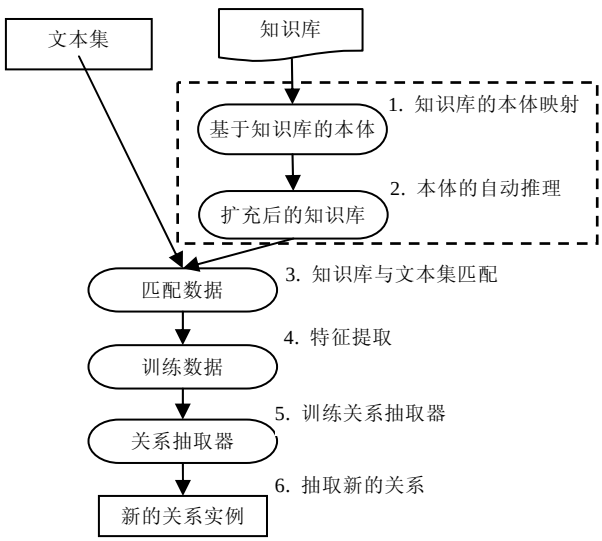

Fig.3 Framework of distant supervision based ontology

图 3 基于本体的远监督抽取架构

文本主要有如下两个任务:

(1) 样本数据经过本体处理,目标的关系实例是否能够增加;

(2) 定义何种指标计算得到的本体,能够对训练得到的分类器产生最为显著的影响.

针对任务(1),我们随机地选取知识库中的部分关系及其实例作为初始的样本数据λ,通过本体推理,得到数据λ',检查λ与λ'中各个关系实例的个数,比较在推理前后关系实例是否增加以及增加的幅度大小.

针对任务(2),我们定义了 3 种不同的映射评价指标,找到使各个指标取最大值的本体文件.初始的样本分别经过各个本体进行处理,提供训练数据.比较不同训练数据得到的分类器的效果,得出结论.

4 实验评测

为了验证本文方法的有效性,我们设计了相关的实验来解决如下几个问题:

(1) 基于本体的样本扩充方法通过不同的标准获取的本体,对于样本关系实例增加的效果;

(2) 基于本体的样本扩充方法通过本体扩充样本数据,与传统的远监督方法相比,最终的关系抽取器的效果是否能够提高;

(3) 哪种指标映射得到的本体,对关系抽取器的影响最大.

4.1 语料库与相关工具

4.1.1 知识库

在本文中,基于本体的样本扩充方法选取 Freebase[19]作为知识库来提供初始的样本数据,Freebase 包含了成千上万种关系以及个体,覆盖了很多的领域.我们选取了不同领域的 6 组样本数据,每组包含了 10 种关系类型,并将各种关系的全部实例作为样本数据.将上述样本数据中的关系实例平均分成两份:一份用在训练阶段,另一部分在测试阶段使用.值得注意的是:由于 Freebase 中的关系名称都是以类似于/location/country/capital 这种形式给出的,这样在利用指标 Cover 时会遇到难以覆盖的问题,所以本文在实验中只截取每个关系名称的最后一部分(即 capital)作为该关系的名称.

4.1.2 文本集

本文使用了维基百科的文本(即 Freebase Wikipedia Extraction,简称 WEX)[20],具体采用的是 2011 年 11 月 11 日导出的数据,它包含了 570 万文章(6 000 万句子).我们选用 Wikipedia 是基于如下考虑:

1) 维基百科是在线更新数据,可以确保所获取数据时效性;
2) Freebase 中的很多关系实例是知识库维护志愿者们从维基百科中手工抽取的,这样,我们在数据对齐的时候就可能找到尽量多的关系实例对应的句子;
3) 维基百科的数据质量比较高,结构性较强,描述的实体关系也较为清晰.

对于获取的维基百科的数据,我们进行预处理,去除那些用户界面内容的文本以及只有讨论内容的文本.对这个文本库进行了分句、词性标注[21]、命名实体识别[22]、依赖解析处理[23],使用的工具是斯坦福自然语言处理组的工具包(Stanford CoreNLP).

4.1.3 相关工具

在本体发现阶段,为了能够减少本实验的计算量,本文借助了本体搜索工具 Swoogle[24]产生候选本体集合.Swoogle 是 Li 等人在 2004 开发的一种基于网络爬虫的语义网检索系统.它抽取每一个已知文件的元数据,并且计算这些文件之间的关系(例如 IM,EX,PV 等).这些已知文件可以通过一个信息检索系统进行查询.这个检索系统是利用特征 *N*-Gram 或者 URIrefs 作为关键字来找到相关的文件,并且可以计算与文件集合的相似性.Swoogle 内部采用一种类似 Page Rank 算法的本体排序算法,将网络上链接的本体文件进行排序.在基于本体的样本扩充方法中,我们通过输入样本中目标关系的领域关键字(比如研究 Location 方面的关系,就利用关键词 Location 进行搜索),Swoogle 会给出与关键字较为匹配结果.我们筛选出前 50 个本体,利用 Pellet 进行本体一致性检测,过滤掉不一致的本体.对于通过一致性检测的本体,再利用基于本体的样本扩充方法定义的不同指标,去寻找相应的指标取最大值的本体.

在本体的自动推理过程,本文采用本体推理工具 Pellet[25].Pellet 是基于表算法的推理机,它会检查知识库的一致性,并且将其他所有的推理服务转化为一致性的检查.我们运用 Pellet 的 ABox 实例查询的功能来寻找样本知识库中实体间的隐含关系类型,并添加到该关系的实例中去.以此就可以得到扩充后的知识库的数据.

4.2 评价标准与实验设计

经过远监督方法训练得到的关系抽取器,我们采用类似于全监督方法的评价标准进行测评,即准确率(precision)和召回率(recall).

为验证实验部分开始提出的 3 个问题以及基于本体的样本扩充方法的有效性,本文设计了如下两组实验:

实验 1. 针对每组的样本数据,我们根据样本映射的 3 种不同的指标(cover,volume,joint),得到 3 个本体文件,经过本体的自动推理以后,比较不同方法的样本数据增加的数量.

实验 2. 我们利用实验 1 中的 3 种指标扩充后的样本数据和初始的样本数据进行启发式的匹配,然后使用远监督的方法去训练关系抽取器,利用准确率和召回率来评测关系抽取器的效果.在关系抽取器方面,与 Mintz 的方法一样,本文使用的是基于高斯回归 L-BFGS 优化的多阶逻辑分类器(http://nlp.stanford.edu/software/classifier.shtml).

4.3 本体发现方法的评测

利用知识库中的6组样本,对于每组的数据,人为地定义该组的关键字,利用Swoogle根据关键字搜索出100个本体,我们筛选出排名前 50 的本体作为候选本体.然后分别使用本体映射的 3 种不同的指标(cover,volume,joint),计算得到相应指标取最大值的本体,并利用该本体进行自动推理.

经过实验,针对每组初始关系实例的数量以及 3 种方法,推理得到扩充后样本中关系实例的个数.图 4 所示最终实验得出的数据.图 4 表明了不同本体映射策略下得到的本体对于样本扩充的影响:一般情况下,使用指标Joint获取的本体,对于样本规模的扩充效果最好,能比先前扩大50%左右;而单独使用指标Cover和指标Volume扩充效果不是很稳定,无法判断哪个指标更好.我们观察样本扩充前后发现,本文中的样本与极小样本集并无必然联系.某些待抽取关系即使在扩充以后也未能达到极小样本集的规模,但有些能够达到.

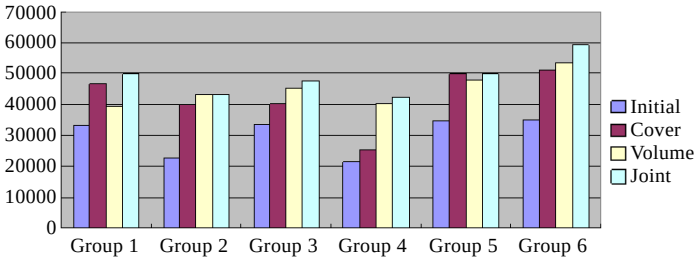

Fig.4 Number of expansion of relation instances

图 4 关系实例的扩充数量

4.4 基于本体的远监督关系抽取任务评测

本节比较5种情况下的关系抽取结果:(1) Mintz的方法;(2) 扩实例较少的本体发现方法(cover,volume); (3) 扩实例最多的本体发现发方法(joint);(4) 由 Hoffmann 等人提出的 MultiR 方法(MultiR)[26];(5) Joint 与 MultiR 相结合的方法(Joint+MultiR).

与 Mintz 方法的实验一样,我们进行了自动评测和人工评测:

- 对于自动测评,我们在训练之前预留出部分的关系实例,在完成抽取器的训练以后进行测试,如果抽取器新发现的关系实例在预留的那部分,则认为是正确的;否则认为是错误的.这种方法的好处就是不需要人工的介入,完全通过算法实现,可以通过调整参数来达到最佳的效果.但是也有些实例是本身正确的,却没有出现在预留的部分,被算法误判为错误的;
- 而人工测评则是需要请志愿者们人工去判断并标记新抽取出来的实例是否具有目标关系,可以减少算法误判错误.

4.4.1 自动测评

自动评测阶段,将初始的每种关系的实例分为两部分,一部分用来训练关系抽取器,另一部分用在测试阶段.用于训练的关系实例,分别采用 3 种不同的指标去寻找各自映射本体,进行样本数据的扩充.完成扩充后,将用于训练的关系实例与训练文本集进行启发式匹配.然后,我们用训练后的抽取系统,对测试文本集进行关系抽取,如果新抽取出来的关系在用来测试的知识库中,则认为是正确的;否则,认为是错误的.

图 5 显示的是最终实验得出的数据,横坐标是召回率,纵坐标是准确率.我们是参照在召回率取不同的值的时候,获取各个抽取系统的准确率,然后在图中标出.

从图 5 中,可以清晰地比较传统的远监督方法和扩充样本后的方法.我们采用在相同召回率的条件下,比较抽取器的准确率,可以得出如下两个结论:

(1) 经过本体推理后的样本数据,训练得到抽取器的效果要比先前的方法好;

(2) 经过指标 Joint 映射的本体,训练得到抽取器的效果要明显的优于其他两种指标得到抽取器.因此,指

标 Joint 是最适合用来寻找某个本体去扩充样本的规模,达到抽取器精度提高的效果.

散点图的横坐标在 0~0.015 区间时,纵坐标的准确率会出现急剧下降的情况.经过认真的数据分析发现,产生这种情况的原因是受到了假否定(false negatives)数据的影响.假否定数据是指那些实际是正确,但却没有出现在用来测试的知识库中,在计算准确率的时候,被当成了系统抽取的错误数据.

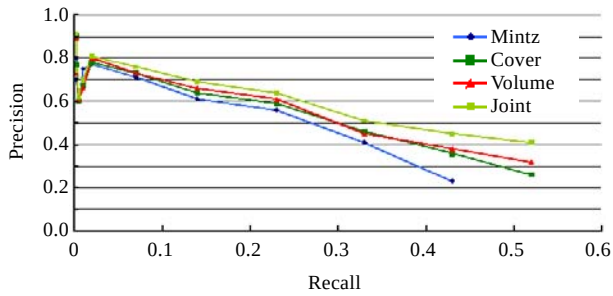

Fig.5 Comparison 1 of relation extraction rate (automatic test)

图 5 关系抽取率对比 1(自动评测)

为了测试本文方法的效果,在上述实验的基础上,我们选取了 Hoffmann 等人在 2011 年提出的 MultiR 方法(先进的多实例关系抽取系统,http://cs.uw.edu/homes/raphaelh/mr)进行了对比实验.MultiR 方法针对远监督学习中出现关系实例重叠的情况做出了改进工作,是目前较为先进的远监督方法.如图 6 所示,我们比较了 Mintz 方法、Joint 方法、MultiR 方法以及 Joint 和 MultiR 相结合的 Joint+MultiR 方法.

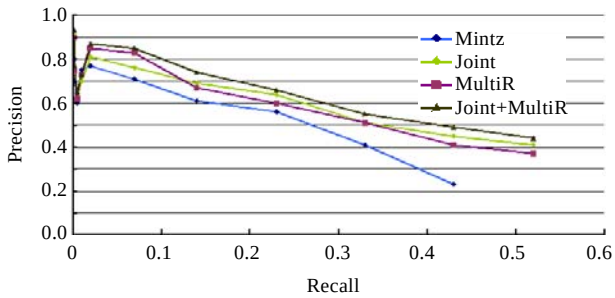

Fig.6 Comparison 2 of relation extraction rate (automatic test)

图 6 关系抽取率对比 2(自动评测)

从图 6 中可以看出,本文的 Joint 方法与 Hoffmann 的 MultiR 方法都可以一定程度上提高分类器的精度.由于 Joint 方法和 MultiR 方法是从不同的角度对于 Mintz 的远监督方法进行改进,所以无法比较两种方法孰优孰劣.但是我们可以比较清晰地看出:两种方法相结合时(即 Joint+MultiR),最终训练得到分类器的效果最好.

4.4.2 人工测评

在人工测试阶段,我们手工判断抽取系统抽出的关系实例是否正确.由于关系种类众多,本部分实验只截取部分关系,见表 1.我们对于分类器抽取的每个关系的前 1 000 个实例进行分析,判断实例的正确性,并且计算得出准确率.

表 1 的实验数据表明:

(1) Joint+MultiR 训练得到的分类器的抽取效果最好;

(2) 经过指标 Joint 得到的本体,将样本扩充后训练得到的抽取系统的准确率普遍比另外两种指标以及 Mintz 的方法要高.

当然,其中也存在某些关系在本体推理前后,抽取系统的表现并没有变好,有些甚至会出现抽取的效果更差

了,表 1 中的关系/location/location/capital 和关系/music/artist/origin.对于表现异常的数据,我们推测有以下两点原因:

- 在本体构建和自动推理的前后,这些关系的实例并没有达到增添的效果,所以训练的样本集也没有扩充,训练数据也就没有改变;
- 这些关系的实例被全部扩充到其他的关系中去了,例如,关系/location/location/capital 被关系/location/location/contains 包含,所以前者的实例都满足后者.在特征抽取的时候,前者那些句子特征也会用于后者.这样,最后训练得到的分类器,对于这两个关系的特征区别将会弱化,所以特征数据的增加反而对分类器分类效果产生了负面的影响.

Table 1 Comparison of relation extraction rate (manual test)

表 1 关系抽取率对比(手工评测)

Relation name	1 000 instances					
	Mintz	Cover	Volume	Joint	MultiR	Joint+MultiR
/location/location/capital	0.69	0.66	0.65	0.63	0.70	0.75
/location/location/containedby	0.75	0.81	0.77	0.81	0.77	0.84
/location/location/contains	0.84	0.86	0.86	0.89	0.88	0.89
/location/location/partially contains	0.57	0.58	0.64	0.65	0.69	0.71
/location/location/partially containedby	0.54	0.61	0.67	0.69	0.70	0.78
/music/artist/origin	0.60	0.58	0.59	0.59	0.57	0.60
/people/person/place of birth	0.78	0.81	0.83	0.87	0.84	0.87
/people/person/parents	0.68	0.69	0.69	0.73	0.72	0.77
/people/person/children	0.47	0.53	0.55	0.61	0.69	0.69
Average	0.65	0.68	0.69	0.72	0.73	0.77

5 结论与展望

基于本体的样本扩充方法是一种新型的远监督学习方法,它对于样本中目标关系实例不足问题有良好的提升效果.实验结果表明:该方法能够有效增加样本的关系实例,尤其是采用 Joint 策略进行本体映射和选取的本体,推理后的样本规模比原先的样本规模扩大了 50%左右.进一步地,利用经过本体扩充后的样本数据进行训练,得到的关系抽取结果要明显优于 Mintz 的远监督学习方法.尤其是 Joint 策略扩展的样本集,抽取器精度提升最为显著,比没有采用本体扩充样本集的普通远监督学习方法高出 6 个百分点左右.

未来的研究工作将继续在如下方面展开:

(1) 探求新的角度来提高远监督方法得到的抽取器的精度;

(2) 将本体的自动推理方法运用到其它机器学习方法中去,来改善机器学习的效果.

致谢 感谢 Dresden University of Technology 的马跃博士和 Rafael Peñaloza Nyssen 博士对本工作的帮助.

References:

[1] Hoffmann R, Zhang C, Weld DS. Learning 5 000 relational extractors. In: Proc. of the 48th Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2010. 286−295.

[2] Carlson A, Betteridge J, Wang R C, Hruschka Jr ER, Mitchell TM. Coupled semi-supervised learning for information extraction. In: Proc. of the third ACM Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2010. 101−110. [doi: 10.1145/1718487.1718501]

[3] Min B, Shi S, Shi S, Grishman R, Lin CY. Towards large-scale unsupervised relation extraction from the Web. Int'l Journal on Semantic Web and Information Systems (IJSWIS), 2012,8(3):1−23. [doi: 10.4018/jswis.2012070101]

[4] Li GJ, Cheng XQ. Research status and scientific thinking of big data. Bulletin of Chinese Academy of Sciences, 2012,27(6): 647−657 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-3045.2012.06.001]

[5] Qin XP, Wang HJ, Li FR, Li CP, Chen H, Zhou X, Du XY, Wang S. New landscape of data management technologies. Ruan Jian Xue Bao/Journal of Software, 2013,24(2):175−197 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4345.htm [doi: 10.3724/SP.J.1001.2013.04345]

[6] Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In: Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int'l Joint Conf. on Natural Language Processing of the AFNLP. Morristown: Association for Computational Linguistics, 2009. 1003−1011.

[7] Wu F, Weld DS. Autonomously semantifying wikipedia. In: Proc. of the 16th ACM Conf. on Information and Knowledge Management. Lisboa: ACM Press, 2007. 41−50. [doi: 10.1145/1321440.1321449]

[8] Li JZ, Liu MX. An important aspect of big data: Data usability. Journal of Computer Research and Development, 2013,50(6): 1147−1162 (in Chinese with English abstract).

[9] Nguyen TVT, Moschitti A. End-to-End relation extraction using distant supervision from external semantic repositories. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Morristown: Association for Computational Linguistics, 2011. 277−282.

[10] Takamatsu S, Sato I, Nakagawa H. Reducing wrong labels in distant supervision for relation extraction. In: Proc. of the 50th Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2012. 721−729.

[11] Bach N, Badaskar S. A review of relation extraction. Literature Review for Language and Statistics II, 2007. 1−15.

[12] Zhou GD, Zhang M, Ji DH, Zhu QM. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In: Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Morristown: Association for Computational Linguistics, 2007. 728−736.

[13] Snow R, Jurafsky D, Ng AY. Learning syntactic patterns for automatic hypernym discovery. In: Proc. of the Advances in Neural Information Processing Systems 17. Vancouver: MIT Press, 2004. 1297−1304.

[14] Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. In: Proc. of the 7th Int'l Conf. on Intelligent Systems for Molecular Biology. Heidelberg: AAAI, 1999. 77−86.

[15] Baader F. The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge: Cambridge University Press, 2003.

[16] Ye YX, Ouyang DT. New research advances in technologies of semantic Web search. Chinese Journal of Computers, 2010,37(1): 1−5 (in Chinese with English abstract).

[17] Lin D. Automatic retrieval and clustering of similar words. In: Proc. of the 17th Int'l Conf. on Computational linguistics. Vol.2. Morristown: Association for Computational Linguistics, 1998. 768−774. [doi: 10.3115/980691.980696]

[18] Horrocks I, Sattler U. A tableau decision procedure for SHOIQ. Journal of Automated Reasoning, 2007,39(3):249−276. [doi: 10.1007/s10817-007-9079-9]

[19] Bollacker KD, Cook RP, Tufts P. Freebase: A shared database of structured general human knowledge. In: Proc. of the 22nd AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2007. 1962−1963.

[20] Google. Freebase Wikipedia extraction (WEX). 2011.

[21] Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proc. of the 43rd Annual Meeting on Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2005. 363−370. [doi: 10.3115/1219840.1219885]

[22] Klein D, Manning CD. Accurate unlexicalized parsing. In: Proc. of the 41st Annual Meeting on Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2005. 423−430. [doi: 10.3115/1075096.1075150]

[23] Bunescu RC, Mooney RJ. A shortest path dependency kernel for relation extraction. In: Proc. of the Conf. on Human Language Technology and Empirical Methods in Natural Language Processing. Morristown: Association for Computational Linguistics, 2005. 724−731. [doi: 10.3115/1220575.1220666]

[24] Ding L, Finin T, Joshi A, Pan R, Cost RS, Peng Y, Reddivari P, Doshi V, Sachs J. Swoogle: A search and metadata engine for the semantic Web. In: Proc. of the 13th ACM Int'l Conf. on Information and Knowledge Management. Washington: ACM Press, 2004. 652−659. [doi: 10.1145/1031171.1031289]

[25] Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y. Pellet: A practical owl-dl reasoner. Web Semantics: Science, Services and Agents on the World Wide Web, 2007,5(2):51−53. [doi: 10.1016/j.websem.2007.03.004]

[26] Hoffmann R, Zhang C, Ling X, Zettlemoyer L, Weld DS. Knowledge-Based weak supervision for information extraction of overlapping relations. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol.1. Morristown: Association for Computational Linguistics, 2011. 541−550.

附中文参考文献:

[4] 李国杰,程学旗.大数据研究:未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考.中国科学院院刊,2012,27(6):647−657. [doi: 10.3969/j.issn.1000-3045.2012.06.001]

[5] 覃雄派,王会举,李芙蓉,李翠平,陈红,周烜,杜小勇,王珊.数据管理技术的新格局.软件学报,2013,24(2):175−197. http://www.jos.org.cn/1000-9825/4345.htm [doi: 10.3724/SP.J.1001.2013.04345]

[8] 李建中,刘显敏.大数据的一个重要方面:数据可用性.计算机研究与发展,2013,50(6):1147−1162.

[16] 叶育鑫,欧阳丹彤.语义 Web 搜索技术研究进展.计算机科学,2010,37(1):1−5.

欧阳丹彤(1968－),女,吉林长春人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为人工智能,自动推理.

E-mail: ouyd@jlu.edu.cn

瞿剑峰(1990－),男,硕士生,主要研究领域为机器学习,语义 Web.

E-mail: qujf13@mails.jlu.edu.cn

叶育鑫(1981－),男,博士,讲师,主要研究领域为逻辑推理,本体论.

E-mail: yeyx@jlu.edu.cn