DOI:10.10.3969/j.issn.1673-4785.201311017
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20150113.1130.008.html

基于弱监督学习的中文网络百科关系抽取

贾真,何大可,杨燕,杨宇飞,冶忠林
(西南交通大学 信息科学与技术学院,四川 成都 610031)

摘 要:实体关系抽取在信息检索、自动问答、本体学习等领域都具有重要作用。提出了基于弱监督学习的关系抽取框架。首先利用知识库中已有结构化的关系三元组,从自然语言文本中自动获取训练语料;针对训练语料数量较少导致特征不足的问题,采用基于朴素贝叶斯的句子分类器和基于自扩展的训练方法,从未标注数据中获取更多的训练语料;然后利用条件随机场模型训练关系抽取器。实验结果表明所提方法的有效性,有现有方法相比,文中方法获得较高的准确率。

关键词:知识获取;信息抽取;关系抽取;弱监督学习;自扩展;中文网络百科;条件随机场;朴素贝叶斯
中图分类号: TP391 **文献标志码**:A **文章编号**:1673-4785(2015)01-0113-07

中文引用格式:贾真,何大可,杨燕,等.基于弱监督学习的中文网络百科关系抽取. 智能系统学报, 2015, 10(1): 113-119.
英文引用格式:**JIA Zhen,HE Dake,YANG Yan,et al.Relation extraction from Chinese online encyclopedia based on weakly supervised learnin[J]. CAAI Transactions on Intelligent Systems**, 2015, 1(6): 113-119.

Relation extraction from Chinese online encyclopedia based on weakly supervised learnin

JIA Zhen, HE Dake, YANG Yan, YANG Yufei, YE Zhonglin
(School of Information and Science Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract: Entity relation extraction plays an important role in the fields of information retrieval, automatic question answering and ontology learning. An entity relation extraction frame based on weakly-supervised learning is proposed in the paper. First, training data are acquired automatically from natural language texts by using relation triples in structured knowledge base. To solve the problem that the number of training data is small and features are insufficient, a bootstrapping method is used to train sentence classifiers based on naive Bayes model. This method can acquire more training data from unlabelled data. The relation extractors are trained by using conditional random fields (CRF) model. The experiment results showed that the method is feasible and effective. Compared with the existing methods state-of-the-art method, the proposed method achieves high accuracy.

Keywords: knowledge acquisition; information extraction; relation extraction; weakly supervised learning; bootstrapping; Chinese online encyclopedia; conditional random fields; naive Bayes

实体关系抽取是自动构建知识库的基础,同时在自动问答、信息检索等多个领域具有重要的应用价值。传统实体关系抽取方法主要有基于模式匹配或基于有监督的统计机器学习。随着关系抽取从限定关系类型转向开放领域,数据源从标准语料库转向海量的网络数据,传统基于模式匹配和有监督统计机器学习的方法逐渐显示出局限性。由于开放领域的关系类型数量巨大,不同关系的模式表现形式多样,变化较大,在基于模式匹配的方法中,难以用人工方式定义全部的模式。在基于有监督机器学习方法中,人工标注训练语料需要耗费大量的人力和时间,面向海量的网络数据,人工标注几乎是不可能

收稿日期:2013-11-07. **网络出版日期**:2015-01-13.
基金项目:国家自然科学基金资助项目(61170111, 61134002, 61202043, 61262058).
通信作者:贾真.E-mail:zjia@home.swjtu.edu.cn.

的。如何能够监督最小化,即不使用人工标注或减少人工标注,也能构建高性能的关系抽取系统是当前的研究热点。由于基于弱监督学习(weakly supervised learning)的关系抽取方法能够在较少人工干预下、自动获取训练语料而受到了广泛的关注。基于弱监督学习的关系抽取框架依赖于一个某领域的知识库,从知识库中可以获取关系三元组,同时需要大量的文本集。从文本集中寻找含有关系实体对的句子,用来建立训练集,然后用这个训练集训练抽取器,从测试文本集中抽取关系实例。现有方法都是自动抽取含有关系实体对的句子作为训练语料,这种利用实体对共现得到的训练语料很不可靠,例如,从知识库中获取关系三元组,〈鲁迅,国籍,中国〉,从文本集中获取含有实体对〈鲁迅,中国〉的句子:"鲁迅以小说创作起家。1918年在《新青年》杂志发表的《狂人日记》是中国现代白话小说的开山之作,影响深远"。这句话并没有表达鲁迅国籍是中国的关系。Riedel 等[1]在纽约时报文本集中进行统计,发现含有国籍关系实体对的句子中38%的句子没有表达国籍这个关系,含有出生地关系实体对的句子中有35%的句子没有表达出生地关系。利用有噪声的训练语料训练模型会影响准确率,降低抽取性能。为了提高训练语料的准确率,文中利用关系词语对训练语料进行约束,即句子中不仅要有实体对,还要有表达关系的词语。由于知识库中的关系实例数量有限,导致训练语料可能较少,存在特征不足的问题,文中利用训练语料训练句子分类器,并基于 bootstrapping 方法迭代地从未标注数据中获取新的训练语料。最后利用 CRF 模型训练关系抽取器。文中的主要贡献有:

1)与利用实体对获取训练语料相比,利用关系三元组获取训练语料的质量有了明显提升;

2)引入了句子分类器从未标注语料中提取新的训练语料,缓解了训练语料不足问题;

3)以互动百科信息盒中的关系实例作为知识库,互动百科条目文本作为训练文本集和测试文本集进行实验,验证了文中方法的有效性。

1 相关工作

实体关系抽取研究始于信息理解会议(message understanding conference, MUC)。1998年最后一次MUC-7上首次提出了关系抽取任务。在MUC-7之后,MUC被自动内容抽取(automatic content extraction, ACE)评测所取代。ACE由美国国家标准技术研究院NIST组织,从1999年至2008年已经举办过9次评测,2008年ACE评测改名为文本分析会议(text analysis conference,TAC),从2008年至今已经举行了6次评测。ACE评测中关系抽取任务包括7个大类关系和若干个子关系。实体关系抽取方法主要有模式匹配的方法和机器学习的方法。在模式匹配的方法中,模式的自动获取技术是研究的关键。机器学习方法根据是否需要人工标注训练语料分为有监督机器学习、半监督机器学习和无监督机器学习。有监督学习方法有特征向量的方法[2-4]和核函数的方法[5-6]。半监督学习方法以少量的关系实例为种子,采用不断迭代的方法从未标注语料中抽取可靠性较高的关系实例[7]。无监督关系抽取主要使用聚类方法[8-9],并为聚类后的簇赋予关系名称。

弱监督学习的关系抽取最早由 Craven 和 Kumlien 提出[10],用于从学术文献的摘要中抽取蛋白质与基因之间的关系。Wu 等[11]利用维基百科信息盒中结构化的〈属性,属性值〉二元组对维基百科条目文本的句子进行回标,自动获取属性关系抽取训练语料,并使用 CRF 模型为每个属性训练抽取器。Bunescu 等[12]分别将具有关系的实体对正例和反例作为查询请求,从搜索引擎查询结果中提取包含实体对的句子作为训练语料。Mintz 等[13]从 Freebase www.freebase.com 中获取具有关系的实体对,从维基百科条目文本中获取关系抽取的训练数据。Mintz 的方法基于以下假设:如果2个实体之间存在某种关系,那么所有含有实体对的句子都描述了这个关系。Yao 等[14]对 Mintz 等[13]提出的方法进行了改进,把关系抽取和实体的种类综合考虑,利用实体的类别来过滤掉部分错误的关系。Riedel 等[1]认为 Mintz 的假设过于严格,含有关系实体对的句子并不一定表达了该关系。Riedel 将 Mintz 的假设放松为:如果2个实体之间存在某种关系,那么含有实体对的句子中至少有一个句子描述了该关系。Surdeanu 等[15]基于弱监督学习对 TAC-KBP 进行属性模板填充,先将维基百科信息盒中的半结构化信息映射至 KBP 结构化的属性模板,再从语料中获取包含实例名和属性值二元组的句子作为训练语料。陈立玮和冯岩松等[16]从互动百科信息盒中获取实体对,从新闻数据中获取训练语料,提出了 bootstrapping 思想的协同训练方法来对弱监督关系抽取模型进行强化,并提出了将传统特征与 *n*-gram 特征相结合进行协同训练的方法。

2 弱监督学习的关系抽取方法

弱监督学习的关系抽取框架包括3个重要的因

素:知识库、训练语料和抽取模型。

2.1　知识库

互动百科是目前最大的中文网络百科之一,互动百科的部分条目中,存在人工创建的信息盒,信息盒中包含了大量半结构化的关系三元组。例如,从互动百科条目“西南交通大学”信息盒中能够获取关系三元组〈西南交通大学,创建时间,1896 年〉。其中,“西南交通大学”是关系主体,“1896 年”是关系客体,创建时间为关系词语。经统计发现,互动百科信息盒中的关系名称是统一定义的,具有较好的唯一性和标识性。文中利用互动百科信息盒获取关系三元组,构造知识库。信息盒是半结构化信息,某些关系有多个客体(如“知名校友”一栏中有多个人名),某些关系的客体不是实体词,而是 1 个句子(如“校训”),因此需要对其进行结构化处理。由于实体关系抽取依赖于命名实体识别,因此只提取主体和客体是命名实体的关系。对于 1 个关系具有多个客体的情况,例如“知名校友”信息盒中的内容为:林同炎,刘大同,张维。分别组成 3 个关系三元组:〈西南交通大学,知名校友,刘大同〉、〈西南交通大学,知名校友,林同炎〉和〈西南交通大学,知名校友,张维〉。

2.2　训练语料

2.2.1　训练语料自动获取

现有弱监督学习的关系抽取框架是利用关系实体对从文本中获取训练语料的,然而,包含实体对的句子有时并不是关系描述语句。基于实体对的共现自动建立起来的训练语料中有大量的噪声,正确的训练语料并不多。为了提高训练语料的准确率,假设关系描述语句中通常以某个特定的关系关键词为核心,例如,“所属地区”关系的表达可能为“位于、处于、属于”等。“创建时间”关系的表达可能为“创立、创建、成立”等。知识库中的关系名称是统一的,然而语句中对关系的表达有多种方式。如果直接匹配关系词语,会导致过低的召回率,因此,需要将关系关键词进行同义扩展。

首先采用西南交通大学中文分词[17]对关系词语进行细粒度分词。例如,“创建时间”细粒度分词后为“创建”和“时间”2 个词。然后分别对这 2 个词语进行同义扩展。文中采用基于同义词词林[18]的语义相似度计算扩展关系词语。语义相似度计算采用田久乐等[19]提出的方法。该方法是根据词语的义项在同义词词林的位置和编码计算出词语的语义相似度。满足相似度阈值的词语都视为同义词。词语相似度的值受到 3 个因素的制约:分支层系数、分支层节点总数和分支间隔。为了避免语义漂移,文中仅计算第 5 层分支词语间的语义相似度。例如,“创建”的同义词为“创立、开创、创始、创建、创办”等,“时间”的同义词为“时间、时刻、时日、工夫、日子、光阴”等。将扩展后的这些同义词组合成为新的关系关键词,例如“创立时间、开创时刻”等。同义词词典的关系关键词获取受到词典规模的限制,无法对未登录词进行同义扩展。因此,在提取训练语料时,若某一含有关系实体对的句子中某词语与关系关键词的字面相似度[20]大于 45%,该词语成为关系关键词,该句子成为训练数据。

从知识库中提取关系三元组〈西南交通大学,创建时间,1896 年〉,百科文本集中有一个包含西南交通大学和 1896 年的句子,该句子同时包含关系关键词“创建”,如图 1 所示。提取该句子作为“创建时间”关系的训练语料。

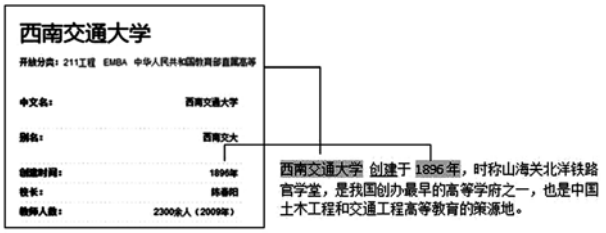

图 1　从文本集中获取训练数据

Fig.1　Training data acquisition from texts

2.2.2　基于句子分类器的训练语料优化

与人工标注的可靠的训练语料不同,由于自然语言预处理错误或知识库中的关系客体在文本中不存在,就会导致错误和遗漏的标注。特别是由于知识库中的关系实例数量较少时,自动获取的训练语料数量较少,许多测试数据中的特征在训练语料中很少出现甚至不存在。文中将已标注的训练语料作为正例,从未标注数据中提取部分数据作为反例,采用 bootstapping 方法训练分类器,然后对未标注数据进行分类,标注为正例的数据作为新的训练数据。

一个分类器性能的优劣往往取决于选择的特征是否能够最大程度地表达不同类别的差异,选择恰当的特征有助于学习到性能较好的分类器,实现不同类别的最优划分。句子分类常用的特征包括词法特征、句法特征和 n -gram 特征。词法特征由句子中的词序列和词性序列构成,而句子中的语言描述过于具体,很难在其他的句子中再次出现,导致严重的数据稀疏性问题,也使得训练出的模型缺乏泛化能力。句法特征从句子的依存句法分析结果中获取。句法特征也存在词法特征中的数据稀疏性问题,并且句法特征依赖于句法分析的效果,然而现有中文

句法分析工具的准确率都不是很理想,导致句法特征不可靠。n-gram 特征通常是文本中 n 个连续词组成的序列,可以捕捉到局部范围内连续词语之间的序列关系,体现语法习惯,n-gram 只包含 3~4 个词,因而不会像传统词法特征那样过于具体,导致特征稀疏,几乎不可能再现。除了传统词语序列的 n-gram 特征,文献[16]把连续词语的词性标注组织成词性序列 n -gram 特征;以及把词语和它的词性序列组成 n-gram 特征,并使用 tri-gram,即$n=3$。

文中采用由词语和它的词性组成的 n-gram 特征,并令 $n=1,2,3$。

1-gram:1 个词语+词性 ($word_i$/ pos_i) (wordi)

2-gram:2 个连续词语+词性($word_i$/ pos_i, $word_{i+1}$/ pos_{i+1})

3-gram:3 个连续词语+词性($word_i$/ pos_i, $word_{i+1}$/ pos_{i+1}, $word_{i+2}$/ pos_{i+2})

从句子中 2 个实体词之间的文本中提取 1/2/3-gram 作为特征值。1/2/3-gram 表示既取 1-gram,又取 2-gram、3-gram。例如句子"英国威尔士大学/ntu 圣三一学院/nt 成立/v 于/p 1848 年/t"中提取了多个 1-gram "圣三一学院 / nt"、"成立 / v"、"于 / p"等,以及多个 2-gram"圣三一学院 / nt , 成立 / v"、"成立 / v , 于 / p"等。

文中利用朴素贝叶斯分类(naïve Bayes classification,NBC)模型训练句子分类器。训练数据作为正例,从未标注数据中提取部分数据(未标注数据中也含有实体对)作为反例,首先提取正例特征和反例特征训练分类器,然后对未标注数据进行分类,对新正例进行标注,并将新正例加入到训练语料中。对新正例进行标注的方法是根据实体类别分别标注关系主体和关系客体,将出现概率最大 n -gram 标注为关系关键词。例如,1-gram"成立"出现概率最大,那么"成立"就是关系关键词,若句子中有多个关系主体或客体,则标注最先出现的实体对作为关系主体和客体。

2.3 抽取模型

条件随机场(conditional random field,CRF)是由 Lafferty 等[21]于 2001 年首先提出,是目前优秀的机器学习模型之一。已被广泛用于中文分词、实体识别、词性标注和信息提取等自然语言处理领域。CRF 是一个判别式模型,其最简单的形式是线性的 CRF,即模型中各个节点之间构成线性结构。一个线性的 CRF 对应于一个有限状态机,它非常适合于进行线性数据序列的标注,在信息提取任务中,基于 CRF 用于序列标注的优势,将信息提取问题转换成目标信息的序列标注问题。

为了进行 CRF 的训练,文中在训练语料中引入标注集对训练语料进行转换。文中使用的是 BIESO 序列标注集,其中 B 代表关系关键词的开始;I 代表关系关键词的内部;E 代表关系关键词的结尾;S 代表独立的实体;O 代表了当前词既不是实体,也不是关系关键词。

例如对训练语句"滨海大学/ntu 是/vshi 一所/mq 综合性/n 国立大学/nis ,/w 成立/vi 于/p 1991 年/t"进行序列标注如图 2。

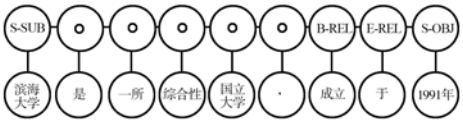

图 2 序列标注示例

Fig.2 Sequence labeling example

图 2 中,S-SUB 表示关系主体,S-OBJ 表示关系客体,B-REL 表示关系关键词的开始,E-REL 表示关系关键词结尾。将该训练语句转换为序列标注的形式如表 1。

表 1 训练语句序列标注

Table 1 Sequence labeling of training sentenc

特征 1	特征 2	标注
滨海大学	ntu	S-SUB
是	shi	O
一所	mq	O
综合性	n	O
国立大学	nis	O
,	w	O
成立	v	B-REL
于	p	E-BEP
1991 年	t	S-OBJ

在 CRF 建立抽取模型过程中,特征选取与特征模板的制定是一项重要任务。特征选取时 CRF 模型不仅能够综合使用字、词、词性、词形等上下文信息,还能利用各种外部信息,如词典等。文中选取词本身、词性、上下文信息(定义活动的窗口)作为特征。"上下文信息"指的是包括当前词在内的及其前后若干个词所组成的观察窗口[22]。窗口过大,选择的特征会急剧增加,影响运行效率;窗口过小,选择的特征较少,影响抽取器的性能。根据分析,选择长度为 3 的窗口,即观察包含当前词在内以及其前后各 2 个词。特征模板如表 2 所示。

表 2　特征模板

Table 2　Feature templates

模板形式	模板含义
W(0)	当前词
W(-1)	当前词左边第 1 个词
W(-2)	当前词左边第 2 个词
W(1)	当前词右边第 1 个词
W(2)	当前词右边第 2 个词
P(0)	当前词词性
P(-1)	当前词左边第 1 个词的词性
P(-2)	当前词左边第 2 个词的词性
P(1)	当前词右边第 1 个词的词性
P(2)	当前词右边第 2 个词的词性
W(0)P(0)	当前词和词性
W(-1)P(-1)	当前词左边第 1 个词和词性
W(-2)P(-2)	当前词左边第 2 个词和词性
W(1)P(1)	当前词右边第 1 个词和词性
W(2)P(2)	当前词右边第 2 个词和词性

3　实验与结果分析

弱监督关系抽取需要结构化知识库构建训练集。知识的表达形式为三元组,即〈主体,关系,客体〉的结构。选用互动百科信息盒构造结构化知识库并进行分词和实体识别预处理,关系三元组中的主体和客体必须为命名实体,从中选取“创建时间”、“国籍”、“出生年月”、“所属地区”等 4 种关系进行实验。文本集采用互动百科条目文章。这里以“创建时间”关系为例对语料进行说明。关系主体类型为机构名(词性标注为 nt),客体类型为时间(词性标注为 t)。从信息盒中抽取关系三元组共有 9 257 个,匹配句子有 6 876 个,从其余未匹配的句子中提取含有实体二元组〈nt, t〉的句子作为测试文本集,测试文本集中的句子数共 114 831 个。

实验结果的评价包括分类器的评价和抽取器的评价。分类器的性能用从测试文本集中得到的正例中正确标记的关系比率来评价,正确的比率越大说明分类器的性能越好。实验从测试文本集中得到正确的正例数量为 T_2,正例总数记为 T_1。准确率 P_C 计算公式如式(1):

$$P_C = \frac{T_1}{T_2} \times 100\% \tag{1}$$

由于从测试文本集中生成的正例总数较多,故采用随机抽样的评价方法。文中设计了 3 种 n-gram 特征:词语序列特征、词性序列特征、以及词语和词性组合序列特征。这里分别测试了 3 种特征的分类器准确率(P_C),测试结果如表 3 所示。

表 3　不同特征下分类器性能比较

Table 3　Performance comparison among different feature　%

特征	创建时间	国籍	所属地区	出生年月
词语序列	72	94	98	96
词性序列	70	84	92	90
词语+词性序列	72	90	94	90

通过表 3 看出,用词语序列作为特征的分类器准确率最好,其次是词语+词性序列特征。然而词语序列作为特征的分类器获取的新正例数量较少,用词语+词性序列特征和词性特征获取新正例的数量均较多。例如表 4 中,对于“创建时间”关系,用词语+词性序列获取的新正例数量为 4 174 个,用词语序列获取的新正例仅为 2 697 个;对于“出生年月”关系,用词语+词性获取的新正例数量为 3 491 个,用词语序列获取的新正例仅有 1 795 个。因此,采用词语+词性组合特征的分类器总体性能最好。

表 4　不同特征下训练语料数量比较

Table 4　quantity comparison among different feature　%

特征	创建时间	国籍	所属地区	出生年月
词语序列	2697	88	4234	1795
词性序列	4174	136	6820	3491
词语+词性序列	4174	136	6820	3491

在抽取器的评价中,文中采用关系的准确率(P)、召回率(R)、F 值(F-Score)作为最终的评价标准,计算方法如式(2)~(4)所示。

$$P = \frac{V_1}{V_2} \times 100\% \tag{2}$$

$$R = \frac{V_1}{V_3} \times 100\% \tag{3}$$

$$F - \text{Score} = \frac{2 \times P \times R}{P + R} \times 100\% \tag{4}$$

式中:V_1是抽取正确的关系个数;V_2是抽取关系的总个数;V_3是语料中关系的个数。

将本中方法与不采用分类器直接利用三元组获取的训练语料训练 CRF 抽取器的抽取结果进行对比,对比结果如表 5 所示。从表 5 可以看出,与未经过训练语料优化而直接采用 CRF 训练抽取器的方法相比,文中方法在保持了较高准确率的基础上,召回率也有了较大的提高。说明利用朴素贝叶斯分类器从反例中获取新正例来优化训练语料,在一定程度上提高了训练语料的质量和抽取的性能。在以上

4种关系抽取中,创建时间关系的准确率和召回率均较低,这是由于句子中的关系主体(类型为nt)或关系客体(类型为t)不唯一,例如大学机构往往有子机构(如院系等),以子机构的创建时间作为关系客体则会造成错误。

表5 与未优化训练语料的关系抽取方法对比

Table 5 Performance comparison with the method of unoptimized training corpus %

关系	文中方法			未优化训练语料		
	准确率	召回率	*F* 值	准确率	召回率	*F* 值
创建时间	72	11.9	20.6	68	8.8	15.6
国籍	84	84.6	84.3	98	72.1	83.1
出生年月	96	27.5	42.8	96	33.4	49.6
所属地区	98	95.3	96.6	98	77.3	86.4

现有弱监督学习的关系抽取框架是将关系抽取看做一个分类问题,首先利用实体对获取训练语料,然后训练分类器,从测试文本集的句子中提取实体对,利用分类器对实体对进行关系预测。文中与文献[13]的方法进行对比,对比结果如表6。

表6 与其他弱监督学习的关系抽取方法对比

Table 6 Performance comparison with other weakly supervised method %

关系	文中方法			文献[13]方法		
	准确率	召回率	*F* 值	准确率	召回率	*F* 值
创建时间	72	11.9	20.6	47	99	55.1
国籍	84	84.6	84.3	0.02	99	0.03
出生年月	96	27.5	42.8	66.6	70.2	68.4
所属地区	98	95.3	96.6	37.7	99	54.6

通过表6的实验结果可以看出,现有弱监督学习的关系抽取系统获得较高的召回率,然而,关系预测的准确率非常低,这是由于没有关系词语的约束会导致关系识别错误。尤其在“国籍”关系抽取中,句子中人名和地名共现的情况非常多,而仅有较少的句子表达国籍关系。文中方法的准确率普遍较高,而且“国籍”关系和“所属地区”关系抽取也取得了较高的召回率,总体抽取性能优于现有弱监督学习的关系抽取方法。此外,对于简单句子的抽取效果较好,复杂句子或长句子的抽取效果不好。分词、词性标注、实体标注等自然语言预处理错误对于关系抽取性能也会产生影响。

4 结束语

文中提出了一种弱监督学习的关系抽取方法框架,该方法从中文网络百科条目半结构化的信息盒中提取关系三元组构建知识库,利用关系三元组对百科文本中进行回标,包含实体对和关系词语的句子成为关系抽取的训练语料,该方法有效解决了训练语料自动构建的问题。针对训练语料较为稀疏从而导致特征不足的问题,提出了bootstrapping的训练语料优化方法,该方法以已标注的训练语料为正例,以部分未标注数据为反例,训练贝叶斯分类器,然后从未标注数据中提取新的正例,补充训练语料的不足。对于分类器特征提取问题,论文提出一种词和词性组合的 *n* -gram 特征,从正例和反例的句子中分别提取词语和词性组合的1/2/3-gram作为特征,训练分类器。实验结果表明优化训练语料能够提升关系抽取的性能。利用关系词语对训练语料和测试语料进行约束,与仅利用实体对共现获取的训练语料进行关系抽取相比,抽取准确率有了显著提高。

参考文献:

[1] RIEDEL S, YAO L, MCCALLUM A. Modeling relations and their mentions without labeled text[J]. Machine Learning and Knowledge Discovery in Databases, 2010, 6323: 148-163.

[2] ZHANG T, Regularized winnow methods[J]. Advance in Neural Information Processing Systems, 2001(13): 703-709.

[3] KAMBHATLA N. Combining lexical, syntactic and semantic features with maximum entropy models for extracting relations[C]//Proceedings of the ACL, 2004 on Interactive Poster and Demonstration Sessions. Barcelona, Spain, 2004: 178-181.

[4] TRATZ S, HOVY E. ISI: automatic classification of relations between nominals using a maximum entropy classifier [C]//Proceedings of the 5th International Workshop on Semantic Evaluation. Uppsala, Sweden, 2010: 222-225.

[5] ZELENKO D, AONE C, RICHARDELLA A. Kernel methods for relation extraction [J]. Machine Learning, 2003(3):1083-1106.

[6] GIULIANO C, LAVELLI A, PIGHIN D, et al. FBK-IRST: Kernel methods for semantic relation extraction[C]//Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007). Prague, Czech, 2007: 141-144.

[7] 程显毅, 朱倩. 未定义类型的关系抽取的半监督学习框架研究[J]. 南京大学学报:自然科学版, 2012, 48(4): 466-474.
CHENG Xianyi, ZHU Qian. A study of relation extraction of undefined relation type based on semi-supervised learning framework[J]. Journal of Nanjing University: Natural Sciences, 2012, 48(4): 466-474.

[8] BOLLEGALA D, MATSUO Y, ISHIZUKA M. Relational duality: unsupervised extraction of semantic relations between entities on the Web[C] //Proceedings of the 19th World Wide Web Conference. New York, 2010: 151-160.

[9] YAN Y, OKACAKI N, MATSUO Y, et al. Unsupervised relation extraction by mining Wikipedia texts using information from the Web[C] //Proceedings of the Joint Conference of the 46th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Singapore, 2009: 1021-1029.

[10] CRAVEN M, KUMLIEN J. Constructing biological knowledge bases by extracting information from text sources [C] //Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology. Palo Alto, CA, 1999: 77-86.

[11] WU F, DANIEL S W. Autonomously semantifying wikipedia[C] //Proceedings of the ACM Sixteenth Conference on Information and Knowledge Management. New York, 2007: 41-50.

[12] BUNESCU R C, MOONEY R J. Learning to extract relations from the web using minimal supervision[C] //Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, 2007: 567-570.

[13] MINTZ M, BILLS S, SNOW R. Distant supervision for relation extraction without labeled data [C] //Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, 2009: 1003-1011.

[14] YAO LM, RIEDEL S, MCAALLUM A. Collective cross document relation extraction without labeled data[C] //Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, 2010: 1013-1023.

[15] SURDANU M, MCCLOSKY D, TIBSHIRANI J, et al. A simple distant supervision approach for the TAC-KBP slot filling task [C] //Proceedings of the Text Analysis Conference 2010-Knowledge Base Population Worksho. [s.l.], 2010:1-5.

[16] 陈立玮, 冯岩松, 赵东岩. 基于弱监督学习的海量网络数据关系抽取[J]. 计算机研究与发展, 2013, 50(9): 1825-1835.
CHEN Liwei, FENG Yansong, ZHAO Dongyan. Extracting relations from the web via weakly supervised learning[J]. Journal of Computer Research and Development, 2013, 50(9): 1825-1835.

[17] 尹红风, 贾真, 李天瑞, 等. 西南交通大学中文分词[EB/OL]. [2012-07-24]. http://ics.swjtu.edu.cn.
YIN Hongfeng, JIA Zhen, LI Tianrui, et al. Southwest Jiaotong University Chinese Segmentation [EB/OL]. [2012-07-24]. http://ics.swjtu.edu.cn.

[18] CHE W X, LI Z H, LIU T. LTP: a Chinese language technology platform[C]//Proceedings of the Coling 2010. [s.l.], 2010: 13-16.

[19] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报: 自然科学版, 2010, 28(6): 602-608.
TIAN Jiule, ZHAO Wei. Words similarity algorithm based on Tongyici Cilin in semantic Web adaptive learning system [J]. Journal of Jilin University: Inform ation Science Edition, 2010, 28(6): 602-608.

[20] 张雪英, 闾国年. 基于字面相似度的地理信息分类体系自动转换方法[J].遥感学报, 2008, 12(3): 433-440.
ZHANG Xueying, LU Guonian. Approach to automatic conversion of geographic information classification schemes [J]. Journal of Remote Sensing, 2008, 12(3): 433-440.

[21] LAFFERTY J, PEREIRA F, MCCALUM A. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of 18th International Conference on Machine Learning. San Francisco: AAAI Press, 2001: 282-289.

[22] 张佳宝. 基于条件随机场的中文命名实体识别研究[D]. 长沙:国防科技大学, 2010:45-59.
ZHANG Jiabao. The research on conditional random fields based Chinese named entity recognition[D]. Changsha: National University of Defense Technology, 2010: 45-59.

作者简介:

贾真, 1975 年生,女,讲师,主要研究方向为内容安全、信息抽取、知识工程。四川省计算机学会大数据专委会委员,中国计算机学会中文信息技术专委会委员。

何大可, 1944 年生,男,教授,博士生导师,中国密码学会副理事长、学术委员会委员,信息安全国家重点实验室第四届学术委员会委员,全国并行计算专业委员会委员,中国电子学会高级会员。主要研究方向为信息安全、内容安全、并行计算。曾获陕西省及国家教委科技进步二等奖、国家自然科学四等奖发表学术论文 240余篇,出版专著 3 部。

杨燕, 1964 年生,女,教授,博士生导师,博士,主要研究方向为数据挖掘、计算智能、集成学习。ACM 成都分部副主席,中国计算机学会人工智能与模式识别专委会委员和理论计算机科学专委会委员,中国人工智能学会机器学习专委会委员和粗糙集与软计算专委会委员。曾获四川省优秀教学成果二等奖,校优秀教学成果一、二等奖,发表学术论文 120 余篇,出版专著 1 部。