DOI:10.3969/j.issn.1671-0673.2016.05.006

实体关系抽取研究综述

刘绍毓,李弼程,郭志刚,王　波,陈　刚

(信息工程大学,河南 郑州 450001)

摘要:实体关系抽取作为信息抽取的核心任务和重要环节,能够实现实体对间语义关系的识别,对句子语义理解及实体语义知识库构建有着重要作用。回顾了实体关系抽取的发展史,总结了有监督实体关系抽取、无监督实体关系抽取、半监督实体关系抽取和开放式实体关系抽取4类方法的原理和代表性研究,并对各类方法进行了详细比较。

关键词:实体关系抽取;有监督方法;无监督方法;半监督方法;开放式实体关系抽取方法

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1671-0673(2016)05-0541-07

Review of Entity Relation Extraction

LIU Shaoyu, LI Bicheng, GUO Zhigang, WANG Bo, CHEN Gang

(Information Engineering University, Zhengzhou 450001, China)

Abstract: As a core task and important part of information extraction, entity relation extraction can realize the identification of the semantic relation between entity pairs and plays an important role in semantic understanding of sentences and the construction of entity knowledge base. This paper first reviews the development history of the entity relation extraction, then makes a summary of supervised relation extraction, unsupervised relation extraction, semi-supervised relation extraction and open relation extraction on principles and representative studies. Finally, this paper gives a detailed comparison of the four methods.

Key words: entity relation extraction; supervised relation extraction; unsupervised relation extraction; semi-supervised relation extraction; open relation extraction method

Web2.0 的兴起和互联网的迅猛发展,改变了人们利用人工整理的方式从书本、报纸、电视等传统社会媒体被动获取知识的模式。面对网络信息爆炸式增长所带来的巨大挑战,人们可以借助搜索引擎主动从海量文本中快速查找所需的资料。然而,基于关键词匹配技术和 PageRank 等网页排序技术的搜索引擎虽然能够在一定程度上满足用户信息获取的需求,但仍存在信息过载、资源迷向等问题,无法完成对文本的深层次理解。用户在使用搜索引擎获得大量相关网页之后,仍需要进行繁琐费时的浏览、筛选工作才能获取自己所需的信息。因此,迫切需要文本深层分析工具为用户提供更精准的服务。

实体关系抽取作为信息抽取领域的重要研究课题[1],其主要目的是抽取句子中已标记实体对之间的语义关系,即在实体识别的基础上确定无结构文本中实体对间的关系类别,并形成结构化的数据以便存储和取用,例如,输入一个带有标记实体的句子"<e1>梁稳根</e2>任<e2>三一集团</e2>董事长,是一名优秀的中国民营企业家。",实体关系抽取系统能自动识别实体"梁稳根"和"三一集团"的关系是雇佣关系。

实体关系抽取作为信息抽取的重要任务,是在实体识别的基础上从非结构化文本中抽取出预先

收稿日期:2015-06-05;**修回日期**:2015-08-30

基金项目:国家 863 计划资助项目(2011AA7032030D)

作者简介:刘绍毓(1987-),男,硕士生,主要研究方向为实体关系抽取,E-mail:tyughvbn88@163.com;

李弼程(1970-),男,教授,博士生导师,主要研究方向为语音信号处理与智能信息处理。

定义的实体关系。该技术突破了传统的必须经过人工阅读、理解的方式来获得语义关系的限制,取而代之的是语义关系的自动查找和抽取。从用户需求层面看,文本分类、文本聚类等技术能从大量的文本集合中筛选或组合出用户所需要的文本或段落。而实体关系抽取则可以从更小粒度的文本句子中挖掘出用户所需要的语义关系信息,给用户提供一项更精细的服务。实体关系抽取的结果可用于构建知识图谱或本体知识库,用户可从中检索和使用所需要的知识。实体关系抽取还能为自动问答系统的构建提供数据支持。当用户向自动问答系统提问时,自动问答系统能从其结构化数据库中快速准确地检索到答案并提供给用户。从理论价值层面看,实体关系抽取技术能为其它自然语言处理技术提供理论支持。实体关系抽取在语义网络标注、篇章理解、机器翻译方面具有重要的研究意义。

1 实体关系抽取的发展历程与评价体系

1.1 实体关系抽取的发展历程

1998年,美国国防高级研究计划委员会(defense advanced research project agency, DARPA)资助的最后一届消息理解会议(message understanding conference, MUC)首次引入了实体关系抽取任务[2]。MUC中的模板关系(template relation)是对实体关系的最早描述。作为实体关系抽取研究的首次评测会议,18家科研单位参与MUC评测。MUC的语料文本来自于飞机失事事件(airplane crashes)和航天器发射事件(rocket missile launches),涉及的实体关系类别有3种:LOCATION_OF、EMPLOYEE_OF和PRODUCT_OF[3]。

1999年,美国国家标准技术研究院(national institute of standards and technology, NIST)组织了自动内容抽取(automatic content extraction, ACE)评测,其中的一项重要评测任务就是实体关系识别[4]。ACE实体关系语料定义了7大类实体,包括人物、组织、设施、处所、地理政治实体、车辆、武器,其中每个大类又分为多个子类。

ACE评测会议至今已举办过8次,其实体关系语料由语言资源联盟(linguistic data consortium, LDC)提供[5],涉及的语种由最初的英文扩展到阿拉伯文、中文、西班牙语等。ACE实体关系语料的来源包括网络上的专线新闻(newswire)、通过自动语音识别得到的广播新闻(broadcast conversations)、通过光学字符识别得到的报纸新闻(newspaper)、新闻组(usenet)、对话性的电视谈话(conversational telephone speech)和网络日志(weblog)。其中,中文语料由国内的哈工大自然语言处理实验室标注,语料文本主要来自广播新闻(40%)、新闻专线(40%)和网络对话(20%)。与MUC相比,ACE的实体关系语料的语种数量和数据规模都有了大幅度的增加。ACE 2008的关系抽取任务共定义了Agent-Artifact、General-Affiliation、Metonymy、Organization-Affiliation、Part-Whole、Person-Social、Physical 7个大类的实体关系,细分为User-Owner-Inventor-Manufacturer、Citizen-Resident-Religion-Ethnicity、Organization-Location等18个子类的实体关系[6],ACE评测会议将实体关系抽取研究推到了一个新的高度。从2009年开始,ACE被归入文本分析会议(text Analysis conference, TAC),成为了Knowledge Base Population任务的主要组成部分[7]。

MUC、ACE评测会议的实体关系抽取涉及的关系类型局限于命名实体(包括人名、地名、组织机构名等)之间的少数几种类型的实体关系,如雇佣关系、地理位置关系、人—社会组织关系等。SemEval(semantic evaluation)是继MUC、ACE后信息抽取领域又一重要评测会议,该会议吸引了大量的院校和研究机构参与测评。SemEval-2007的评测任务4定义了7种普通名词或名词短语之间的实体关系,但其提供的英文语料库规模较小。随后,SemEval-2010的评测任务8对其进行了丰富和完善,将实体关系类型扩充到9种,分别是:Component-Whole、Instrument-Agency、Member-Collection、Cause-Effect、Entity-Destination、Content-Container、Message-Topic、Product-Producer和Entity-Origin。考虑到句子实例中实体对的先后顺序问题,引入"Other"类对不属于前述关系类型的实例进行描述,共生成19种实体关系。SemEval-2010评测引发了普通名词或名词短语间实体关系抽取研究的新高潮[8]。

评测会议的参加者大都将实体关系抽取转化为分类问题进行研究。但是,MUC、ACE、SemEval评测会议发布的实体关系语料都是依靠人工标注的方式得到的,即领域专家首先制定好关系类型体系和标注规则,然后从大规模文本逐个进行判断和筛选。此方法耗费大量的人力,成本较高,且语料的扩充困难。此外,该方法获得的实体关系语料领域覆盖面窄,句子实例形式较为单一。

开放式实体关系抽取有效解决了语料获取困

难的问题。互联网开放领域的维基百科、DBpedia、YAGO 和 Freebase 等背景知识库所蕴含的大量事实型信息为标注语料的获取提供了有效的数据支持;相对于传统的人工标注语料,互联网开放语料的规模更大,面向的领域更宽泛,包含的关系类型更全面。

1.2 实体关系抽取的评价体系

实体关系抽取常采用准确率(precision)、召回率(recall)、F 值来进行评价,其计算表达式如下[9]:

$$precision = \frac{某类被正确分类的关系实例个数}{被判定为某类的关系实例总数},$$

$$recall = \frac{某类被正确分类的关系实例个数}{测试集中某类的关系实例总数},$$

$$F1 = \frac{2 * precision * recall}{precision + recall}。$$

2 实体关系抽取的研究现状

根据对标注数据的依赖程度,实体关系抽取方法可分为有监督学习方法、半监督学习方法、无监督学习方法和开放式抽取方法[10],下面分别详细介绍这些方法的研究现状。

2.1 有监督的实体关系抽取

有监督学习方法是最基本的实体关系抽取方法,其主要思想是在已标注的训练数据的基础上训练机器学习模型,然后对测试数据的关系类型进行识别。有监督学习方法包括有基于规则的方法、基于特征的方法和基于核函数的方法[11]。

基于规则的方法需要根据待处理语料涉及领域的不同,通过人工或机器学习的方法总结归纳出相应的规则或模板,然后采用模板匹配的方法进行实体关系抽取。文献[12]用与实体关系有关的语义信息扩展句法分析树,生成规则来进行实体关系识别。近年来,实体关系抽取研究者构建了多个基于规则的实体关系抽取系统[13-16]。然而,该方法的规则制定依赖于专家知识和人工归纳,代价较高,并且该方法领域移植性差,难以得到广泛使用。

基于特征向量的方法是一种简单、有效的实体关系抽取方法,其主要思想是从关系句子实例的上下文中提取有用信息(包括词法信息、语法信息)作为特征,构造特征向量,通过计算特征向量的相似度来训练实体关系抽取模型。该方法的关键在于寻找类间有区分度的特征,形成多维加权特征向量,然后采用合适的分类器进行分类。

在基于特征向量的英文实体关系抽取研究方面,文献[17]利用实体词、实体类型、引用类型等特征构造特征向量,采用最大熵分类器构建抽取模型,在 ACE RDC 2003 英文语料上的实体关系抽取实验表明,该方法在关系小类上获得的 F 值为 52.8%。文献[18]在文献[17]研究的基础上,分类组织各种特征,形成平面组合核,并采用 SVM 分类器在 ACE RDC 2004 英文语料上进行实体关系抽取,获得了 70.3% 的 F 值。文献[19]挖掘出更多种类的特征,比较了不同特征对实体关系抽取性能的影响,并通过实验表明复杂的特征并不一定比简单的特征更有效。文献[20]在文献[17]研究的基础上对关系实例上下文特征进行了扩展,并以布尔型特征向量表示,采用最大熵分类器进行分类,取得了良好的效果。文献[21]在已有特征的基础上,引入字特征,并采用条件随机场(conditional random fields, CRF)进行医学领域实体间关系的抽取,F 值达到 75% 以上。

在基于特征向量的中文实体关系抽取研究方面,文献[22]在 ACE RDC 2004 中文语料库进行实体关系抽取实验,经过反复测试后发现采用实体类型、实体在句子中出现的顺序(或者包含关系)和实体上下文的词等特征能取得最佳抽取性能。文献[23]将实体关系分为包含关系和非包含关系分别进行特征抽取,并采用 CRF 模型进行实体关系抽取,其在 ACE RDC 2007 上进行关系抽取实验的结果表明此方法能显著提高中文实体关系抽取的性能。文献[24]研究了组合特征对实体关系抽取的影响,其在 ACE RDC 2005 中文语料上进行关系抽取的实验结果表明使用组合特征比未使用组合特征效果好。文献[25]在词法特征、实体原始特征的基础上,融入依存句法关系、核心谓词、语义角色标注等特征,实验结果表明该方法能有效提高实体关系抽取的性能。

基于特征向量的实体关系抽取方法能够取得较好的效果,但无法充分利用实体对上下文的结构信息。为此,人们提出了多种基于核函数的实体关系抽取方法,包括词序列核函数方法、依存树核函数方法、最短路径依存树核函数方法、卷积树核函数方法以及它们的组合核函数方法。部分学者还将基于核函数的方法与基于特征向量的方法结合来进行实体关系抽取,实验结果表明基于核函数和基于特征的实体关系抽取方法可以相互补充。

在基于核函数的英文实体关系抽取研究方面:文献[26]率先在文本浅层解析树的基础上定义了核函数及其计算方式,并将其嵌入到 SVM 分类器来抽取 organization-location 和 person-affiliation

的关系实例,通过与基于特征向量的方法的比较,证明了基于核函数的方法能够开发更好的特征集合进行关系抽取。文献[27]提出了与关系句子实例上下文相关的依存树核函数,该方法通过计算依存树的相似度来进行实体关系抽取,实验表明基于依存树核的方法比基于词特征的核方法 F 值大有提高。文献[28]证明基于最短路径核的实体关系抽取方法优于基于依存树核的方法。文献[29]定义了卷积树核,通过求两棵句法解析树的公共子树的个数来计算两棵句法解析树之间的相似度。文献[30]对英文领域的基于核函数的实体关系抽取方法进行了总结。

在基于核函数的中文实体关系抽取研究方面,文献[31]利用卷积核函数中的字符串序列核进行实体关系抽取,并借用《知网》中的词汇语义相似度计算方法计算中文特征词串的相似度,实验结果表明其 F 值达到了84%,这也说明语义信息能提高中文语义关系抽取系统的性能。文献[32]在ACE RDC 2007语料库上对比了卷积树核函数方法和最短依存树核函数方法的性能,实验结果表明基于卷积树核和特征向量核的组合核方法能够有效地进行中文实体关系抽取。文献[33]在卷积树核方法的基础上,对树结构进行了实体语义信息的扩展,并比较了不同实体语义信息对树结构扩展后的系统性能,在ACE RDC 2005中文语料库上进行的关系大类抽取实验的结果表明了核方法进行语义扩展的有效性。文献[34]将最短路径依存树结构分为与上下文信息无关的树结构和与上下文信息相关的树结构两种,并对与上下文相关的树结构进行扩展。实验表明扩展后的树核方法可以与词特征方法相媲美。文献[35]将基于特征向量的平面核与基于句法分析树的结构核相结合,在ACE RDC 2005的中文语料上进行实体关系抽取,F 值比两个单独核函数方法分别高出4.36%和17.37%。文献[36]将反映特定领域实体语义关系领域知识树融合到实例句的句法树中,有效提高了实体关系抽取的性能。

基于核函数的方法不需要构造特征向量,而是把结构树作为处理对象,通过计算它们之间的相似度来进行实体关系抽取。但该方法训练和测试速度太慢,不适合处理大规模数据。

2.2 无监督的实体关系抽取

无监督实体关系抽取方法无需依赖实体关系标注语料,其实现包括关系实例聚类和关系类型词选择两个过程。首先根据实体对出现的上下文将相似度高的实体对聚为一类,然后选择具有代表性的词语来标记这种关系。

文献[37]在计算语言协会(association for computational linguistics, ACL)会议上率先提出无监督实体关系抽取方法,奠定了无监督实体关系抽取的基础。文献[38]提出了一种多层级聚类的无监督实体关系抽取方法。该方法首先将新闻文本按文章的来源进行初始的分类,然后根据语句的语义结构图在一系列约束的情况下抽取出基础模式聚类的实体,这些实体根据基础模式进行映射形成次生聚类,如此循环往复,每个次生聚类中包含具有相同实体关系的实体对。文献[39]对实体关系上下文的特征进行加权,并采用改进的 K 均值算法进行聚类,在ACE语料上的抽取实验结果表明该方法优于Hasegawa方法。文献[40]将卷积树核运用到无监督的中文实体关系抽取中,采用分层聚类算法进行实体关系抽取。文献[41]提出了基于产生式模型的无监督实体关系抽取框架,实现了医学文本中实体关系的有效抽取。文献[42]研究了开放领域的中文实体无监督关系抽取,采用密度聚类算法,取得了较好的实体关系抽取性能。

无监督实体关系抽取无需预先定义实体关系类型体系,具有领域无关性,在处理大规模开放领域数据时具有其它方法无法比拟的优势,但其聚类阈值难以事先确定,并且目前仍缺乏较客观的评价标准。

2.3 半监督的实体关系抽取

文献[43]率先提出基于Bootstrapping的半监督实体关系抽取方法,该方法从包含关系种子的上下文中总结出实体关系序列模式,然后利用关系序列模式去发现更多的关系种子实例,形成新的关系种子集合。重复上述过程,迭代得到实体关系实例和序列模式。该方法获得的序列模式准确率高,但召回率相对较低。为了提高序列模式的召回率,学者们引入软模式(soft-pattern)的概念,对构成模式的元素进行了泛化处理,在一定程度上提高了序列模式的召回率。文献[44]进行了基于中文种子自扩展的命名实体关系抽取,在对《人民日报》语料库的测试中,F 值达到了81.3%。文献[45]将标注传递算法引入实体关系抽取,提出了基于图模型的半监督实体关系抽取,实验表明该方法优于基于SVM的有监督实体关系抽取方法和基于Bootstrapping的半监督关系抽取的方法。文献[46]在单一模式的实体关系抽取系统Snowball的基础上提出了可以提取多种类型实体关系的架构—MultiSnowball,该系统能解决多个关系类型共享模式的问题。文献[47]指出在基于BootStrapping方法的实体关

系抽取方法中,一个关键的问题就是如何对获取的模式进行过滤,以免将过多的噪声引入迭代过程中而导致"语义漂移"问题。为了解决这个问题,文献[48]提出了协同学习(co-learning)方法,该方法利用两个条件独立的特征集来提供不同且互补的信息,从而减少标注错误。文献[49]提出了类型检查(type checking)方法,利用命名实体识别器检查关系实例中涉及的实体类型是否符合该类关系的参数类型,从而达到减少错误标注实例的目的。

半监督实体关系抽取无需大规模标注语料,只需人工标注少量关系实例,因此适用于缺乏标注语料的实体关系抽取。但是,该方法的缺点是对初始关系种子的质量要求较高,领域迁移时需要重新编写规则或构建高质量的关系种子。另外,该方法的召回率较低。

2.4　开放式实体关系抽取

近年来,相关专家学者提出了开放式实体关系抽取方法,该方法能避免针对特定关系类型人工构建语料库,可以自动完成关系类型发现和关系抽取任务。

开放式实体关系抽取方法的基本假设:若已知两个实体存在某种语义关系,所有包含这两个实体的句子都潜在地表达了它们之间的语义关系。开放式实体关系抽取通过借助外部领域无关的实体知识库(如 DBPedia、YAGO、OpenCyc、FreeBase 或其它领域知识库)将高质量的实体关系实例映射到大规模文本中,根据文本对齐方法从中获得训练数据,然后使用监督学习方法来解决关系抽取问题。但是,此方法获得训练语料存在较多噪声,噪声标注的滤除成为该方法的研究重点。

distant supervision 实体关系抽取方法自提出以来,标注数据去噪问题引起了该领域专家学者的普遍关注。文献[50]最早提出 distant supervision 实体关系抽取方法,利用 freebase 知识库和 wikipedia 文本库自动获取关系抽取训练数据(训练数据获取过程实际上也是数据标注过程),并训练模型以实现关系抽取任务。在标注数据获取过程中,Mintz 等假定所有包含实体对的句子都蕴含了两者间的潜在关系。借助于该假定虽然获取特定关系类型的大量正确标注数据,同时也会引入该关系类型的大量噪声文本,称之为噪声标注。

若将所有的已标注关系类型的句子实例都作为正例,将引入大量噪声标注,进而会影响 distant supervision 关系抽取的性能。因此,在训练关系抽取模型之前,需要对候选标注数据中的噪声标注进行识别。

文献[51-52]指出 distant supervision 的基本假设会生成噪声标注,但都未直接给出解决方法。针对 Mintz 的假设,文献[53]提出了一种更加松弛的假设(expressed-at-least-once):若已知某实体对存在某种实体关系,那么至少有一个包含该实体对的句子潜在地表达了这种实体关系。基于此假设,Riedel 获取了更加准确的标注数据。文献[54]认为 Riedel 提出的松弛假设在多数情况下与 Mintz 的基本假设一致。其经过统计得到:当使用 Wikipedia 文本作为文本语料库时,91.7% 的实体对仅仅出现在一个句子实例中,即包含该实体对的句子就被认为蕴含了指定的实体关系。于是其提出了一种模式相关性的方法用以减小 distant supervision 生成的错误标注的数量,并且该方法未使用以上任意一种假设。文献[55]将判决学习方法与主题模型相结合,用以减少 distant supervision 结果中的噪声数据。该方法极大地改善了抽取事实的排序质量,取得良好的抽取性能。表1是上述4种实体关系类型抽取方法的比较。

表1　实体关系抽取方法对比

实体关系抽取方法	实现方法	人工干预程度	领域移植性	性能提升方法
有监督方法	分类	强	弱	改进规则、特征、核函数
无监督方法	聚类	弱	强	扩展特征,改进聚类算法
半监督方法	分类	中	中	改进模式扩展和噪声过滤方法
开放式方法	分类	弱	强	探索文本库的噪声过滤算法

3　结束语

有监督的实体关系抽取方法虽然准确率高,但依赖于标注语料。无监督实体关系抽取无需依赖标注语料,虽然领域移植性强,适合处理大规模开放领域数据,但其聚类阈值难以事先确定,并且目前仍缺乏较客观的评价标准。半监督实体关系抽取只需人工标注少量关系实例,适用于缺乏标注语料的实体关系抽取,但其实现过程中引入的噪声容易造成语义漂移,且该方法的召回率较低。开放式实体关系抽取可以借助互联网自动完成实体关系类型发现和实体关系抽取任务,具有广阔的发展前景和应用空间,将对自动问答系统构建、本体知识库构建、大数据处理、等领域产生深远的影响。

参考文献:

[1] 陈宇, 郑德权, 赵铁军. 基于 Deep Belief Nets 的中文名实体关系抽取[J]. 软件学报, 2012, 23(10): 2572-2585.

[2] Chinchor N, Marsh E. Muc-7 Information Extraction

Task Definition[C]// Proceeding of the Seventh Message Understanding Conference (MUC-7). 1998: 359-367.

[3] Overview of MUC-7/MET-2. [EB/OL] [2005-05-08]. http://www-nlpir. nist. gov/related _ projects/muc/proceedings/muc_7_proceedings/overview. html.

[4] ACE 2005. The Automatic Content Extraction (ACE) Projects[EB/OL] [2007-01-11]. http://www. ldc. upenn. edu/Projects/ACE/.

[5] 赵琦, 刘建华, 冯浩然. 从ACE会议看信息抽取技术的发展趋势[J]. 现代图书情报技术, 2008, 162(3): 18-23.

[6] Chan Y S, Roth D. Exploiting Background Knowledge for Relation Extraction[C]// Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics. 2010: 152-160.

[7] McNamee P, Dang H T, Simpson H, et al. An Evaluation of Technologies for Knowledge Base Population [C]// Proceedings of the Seventh International Language Resources and Evaluation Conference. 2010: 369-372.

[8] Hendrickx I, Kim S N, Kozareva Z, et al. Semeval-2010 task 8: Multi-way Classification of Semantic Relations between Pairs of Nominals[C]// Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. 2009: 94-99.

[9] 李天颖, 刘璘, 赵德旺, 等. 一种基于依存文法的需求文本策略依赖关系抽取方法[J]. 计算机学报, 2013, 36(1): 54-62.

[10] 张传岩. Web实体活动与实体关系抽取研究[D]. 济南: 山东大学硕士学位论文, 2012.

[11] 王敏. 基于多代理策略的中文实体关系抽取[D]. 大连: 大连理工大学硕士学位论文, 2011.

[12] Miller S, Fox H, Ramshaw L, et al. A Novel Use of Statistical Parsing to Extract Information from Text [C]// Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference. 2000: 226-233.

[13] Aitken J S. Learning Information Extraction Rules: An inductive logic programming approach[C]// Proceedings of European Conference on Artificial Intelligence. 2002: 355-359.

[14] Mcdonald D M, Chen H, Su H, et al. Extracting Gene Pathway Relations using a Hybrid Grammar: the Arizona Relation Parser[J]. Bioinformatics, 2004, 20(18): 3370-3378.

[15] Jayram T S, Krishnamurthy R, Raghavan S, et al. Avatar Information Extraction System [J]. IEEE Data Eng. Bull, 2006, 29(1): 40-48.

[16] Shen W, Doan A H, Naughton J F, et al. Declarative Information Extraction using Datalog with Embedded Extraction Predicates[C]// Proceedings of the 33rd International Conference on Very Large Data Bases. 2007: 1033-1044.

[17] 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. 2004: 22.

[18] Zhao S, Grishman R. Extracting Relations with Integrated Information using Kernel Methods[C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. 2005: 419-426.

[19] Jiang J, Zhai C X. A Systematic Exploration of the Feature Space for Relation Extraction[C]// Proceedings of NAACL HLT. 2007: 113-120.

[20] 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. 2010: 222-225.

[21] Miao Q, Zhang S, Zhang B, et al. Extracting and Visualizing Semantic Relationships from Chinese Biomedical Text[C]// Proceedings of the Pacific Asia Conference on Language. 2012: 99-107.

[22] 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.

[23] 董静, 孙乐, 冯元勇, 等. 中文实体关系抽取中的特征选择研究[J]. 中文信息学报, 2007, 21(4): 80-85.

[24] 黄鑫, 朱巧明, 钱龙华, 等. 基于特征组合的中文实体关系抽取[J]. 微电子学与计算机, 2010, 27(4): 198-200.

[25] 郭喜跃, 何婷婷, 胡小华, 等. 基于句法语义特征的中文实体关系抽取[J]. 中文信息学报, 2014, 28(6): 183-189.

[26] Zelenko D, Aone C, Richardella A. Kernel Methods for Relation Extraction[J]. Journal of Machine Learning Researeh, 2003, 3: 1083-1106.

[27] Culotta A, Sorensen J. Dependency Tree Kernels for Relation Extraction[C]// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. 2004: 423-429.

[28] Buneseu R C, Raymond J M. A Shortest Path Dependency Kernel for Relation Extraction[C]// Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. 2005: 724-731.

[29] Zhang M, Zhang J, Su J, et al. A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features[C]// Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. 2006: 825-832.

[30] Choi S, Lee S, Jung H, et al. An intensive case study on kernel-based relation extraction[J]. Multimedia Tools and Applications, 2014, 71(2): 741-767.

[31] 刘克彬, 李芳, 刘磊. 基于核函数中文关系自动抽取系统的实验[J]. 计算机研究与发展, 2007, 44(8): 1406-1411.

[32] Huang R, Sun L, Feng Y. Study of kernel-based methods for Chinese relation extraction[M]. Springer Berlin Heidelberg, 2008: 598-604.

[33] 奚欢欢, 钱龙华, 周国栋,等. 基于合一句法和实体语义树的中文语义关系抽取[J]. 中文信息学报, 2010: 24(5): 17-23.

[34] Zhou G, Qian L, Fan J. Tree Kernel-based Semantic Relation Extraction with Rich Syntactic and Semantic Information[J]. Information Sciences, 2010, 180(8): 1313-1325.

[35] 李丽双, 党延忠, 张婧,等. 基于组合核的中文实体关系抽取研究[J]. 情报学报, 2012, 31(7): 702-708.

[36] 陈鹏, 郭剑逸, 余正涛,等. 融合领域知识短树核函数的中文领域实体关系抽取[J]. 南京大学学报, 2015,51(1):181-186.

[37] Hasegawa T, Sekine S, Grishman R. Discovering Relations among Named Entities from Large Corpora[C]// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. 2004: 415.

[38] Shinyama Y, Sekine S. Preemptive Information Extraction using Unrestricted Relation Discovery[C]// Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. 2006: 304-311.

[39] 张志田. 无监督实体关系抽取方法研究[D]. 哈尔滨:哈尔滨工业大学, 2007.

[40] 黄晨, 钱龙华, 周国栋,等. 基于卷积树核的无指导中文实体关系抽取研究[J]. 中文信息学报, 2010, 24(4): 11-17.

[41] Rink B, Harabagiu S. A Generative Model for Unsupervised Discovery of Relations and Argument Classes from Clinical Texts[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011: 519-528.

[42] 孙勇亮. 开放领域的中文实体无监督关系抽取[D]. 上海:华东师范大学,2014

[43] Brin S. Extracting patterns and relations from the world wide web[M]. Berlin: Springer Heidelberg, 1999: 172-183.

[44] 何婷婷, 徐超, 李晶,等. 基于种子自扩展的命名实体关系抽取方法[J]. 计算机工程, 2006, 32(21): 183-184.

[45] 陈锦瑞, 姬东鸿. 基于图的半监督关系抽取[J]. 软件学报, 2008, 19(11): 2843-2852.

[46] Liu Xiaojiang, Yu Nenghai. MultiType Web Relation Extraction Based on Bootstrapping[C]// Proceedings of WASE International Conference on Information Engineering. 2010: 2427.

[47] Blohm S, Cimiano P, Stemle E. Harvesting Relations from the Web-quantifiying the Impact of Filtering Functions[C]// Proceedings of the National Conference on Artificial Intelligence. 2007: 1316.

[48] Cvitas A. Relation Extraction from Text Documents [C]// Proceedings of the 34th International Convention. 2011: 1565-1570.

[49] Moens M F. Information Extraction: Algorithms and Prospects in a Retreval Context[R]. Springer, 2006.

[50] Mike Mintz, Steven Bills, Rion Snow, 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. 2009: 1003-1011.

[51] Raphael Hoffmann, Congle Zhang, Daniel S. Weld. Learning 5000 Relational Extractors [C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010: 286-295.

[52] Limin Yao, Sebastian Riedel, Andrew McCallum. Collective Cross-Document Relation Extraction Without Labelled Data[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010: 1013-1023.

[53] Sebastian Riedel, Limin Yao, Andrew McCallum. Modeling Relations and Their Mentions without Labeled Text [C]// Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases. 2010: 148-163.

[54] Shingo Takamatsu, Issei Sato, Hiroshi Nakagawa. Reducing Wrong Labels in Distant Supervision for Relation Extraction[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. 2012: 721-729.

[55] Benjamin Roth, Dietrich Klakow. Combining Generative and Discriminative Model Scores for Distant Supervision [C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 24-29.