实体关系抽取的技术方法综述*

徐 健[1,2,3] 张智雄[1] 吴振新[1]
[1](中国科学院国家科学图书馆 北京 100190)
[2](中国科学院研究生院 北京 100049)
[3](中山大学资讯管理系 广州 510275)

【摘要】对实体关系抽取研究以 MUC 和 ACE 评测为主线的发展进行总结,并指出实体关系抽取任务普遍存在的三个问题是特定领域标引数据集的获取、模式的获取以及共指消解。在对当前关系抽取的相关文献、系统和项目进行分析研究的基础上,将基于非结构化文本的实体关系抽取技术方法归纳为:基于模式匹配的关系抽取、基于词典驱动的关系抽取、基于机器学习的关系抽取、基于 Ontology 的关系抽取以及混合抽取方法,旨在为进一步构建实体关系抽取系统提供良好借鉴。
【关键词】实体关系抽取 信息抽取 关系抽取方法
【分类号】G250.73

Review on Techniques of Entity Relation Extraction

Xu Jian[1,2,3] Zhang Zhixiong[1] Wu Zhenxin[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](Graduate University of the Chinese Academy of Sciences, Beijing 100049, China)
[3](Department of Information Management, Sun Yat-Sen University, Guangzhou 510275, China)

【Abstract】Entity relation extraction is a very important task in text information extraction domain. It first summarizes the development of entity relation extraction related to MUC and ACE, and then points out that main difficulties exist in the process of relation extraction are acquisition of training dataset, acquisition of templates, and co-reference resolution. Based on the analysis of recent related literatures, systems and projects, it concludes the entity relation extraction methods as follows: templates method, lexicon driven method, machine learning method, Ontology driven method, and hybrid method. The analysis of these methods can help to build more efficient entity relation extraction system in further step.
【Keywords】Entity relation extraction Information extraction Relation extraction methods

1 引 言

随着数字资源和万维网上的文本信息的飞速增长,从文本中自动抽取知识的技术越来越被关注。具体而言,知识抽取系统从特定领域的文本文档中识别相关知识要素(通常是预先定义的类型),并将它们以结构化形式存储。知识抽取任务在细节和可靠性上有不同的选择,但一般都包括两个普遍存在并且紧密关联的子任务:实体识别和关系抽取。实体识别通过实体抽取技术抽取各个知识要素。抽取出的知识要素以离散的形式存在,只能反映出文本中包含哪些实体,例如人、机构、地点等,却不能反映出知识要素之间的关系,例如机构与人之间的雇用关系、机构与地点之间的位置关系等,而关系抽取则是要解决这一难题。

收稿日期:2008-06-16
*本文系国家社会科学基金项目"从数字信息资源中实现知识抽取的理论和方法研究"(项目编号:05BTQ006)的研究成果之一。

不同的研究者对关系抽取任务的表述不尽相同。Alexander Schutz等人[1]认为关系抽取是自动识别由一对概念和联系这对概念的关系构成的相关三元组。Sophia Katrenko等人[2]则从关系抽取的基本过程角度对关系抽取进行了界定。他们认为,关系抽取可以看作是具有两个步骤的过程,即:识别存在关系的证据和检查是否存在关系。维基百科[3]对关系抽取的解释是,关系抽取是在自然语言处理过程中抽取文本中实体间命名关系的任务。抽取的实体间关系能够通过各种形式/语言来表达。其中一种对网络上数据进行表达的语言是RDF。能用到这些关系的应用领域包括基因-疾病关系,蛋白质相互作用等。而作为关系抽取权威评测会议的ACE(Automatic Content Extraction)[4]将关系抽取任务表述为:探测和识别文档中特定类型的关系,并对这些抽取出的关系进行规范化表示。这些关系中,有些对实体出现顺序敏感,有些对实体出现顺序不敏感。ACE08[5]在以前的关系抽取评测任务基础上新增加了跨文档关系抽取任务。跨文档关系抽取扩展了文档内关系抽取任务,主要用来发现全局实体间的关系。

关系抽取技术在很多领域具有应用价值。在自动问答系统中,关系抽取自动关联相关问题和答案;在检索系统中,关系抽取使类似于"北京有哪些公司?"这样的语义检索功能的实现成为可能;在本体学习过程中,关系抽取能够发现新的实体间关系来丰富本体结构;在语义网标注任务中,关系抽取能够自动关联语义网知识单元。

2 关系抽取研究的发展

关系抽取研究的发展是以MUC(Message Understanding Conference)评测会议和后来取代MUC的ACE(Automatic Content Extraction)评测会议为主线进行的。在这两个会议上,多种先进的信息抽取方法被提出来,在会议提供的数据平台上进行测试,并组织与会者进行讨论。基本上每年一次的评测会议为关系抽取的发展起到了引导和推动作用。

关系抽取任务最初是由美国国防高级研究计划委员会(Defense Advanced Research Projects Agency,DARPA)资助的MUC会议于1998年MUC-7上首次正式提出[6]。MUC的显著特点并不是会议本身,而在于对信息抽取系统的评测。只有参加信息抽取系统评测的单位才被允许参加MUC会议。在每次MUC会议前,组织者首先向各参加者提供样例消息文本和有关抽取任务的说明,然后各参加者开发能够处理这种消息文本的信息抽取系统。在正式会议前,各参加者运行各自的系统处理给定的测试消息文本集合。由各个系统的输出结果与手工标注的标准结果相对照得到最终的评测结果。最后才是所谓的会议,由参与者交流思想和感受。这种评测驱动的会议模式被证明是行之有效的,对信息抽取的发展起到了推动作用。MUC主要关注于包括自由文本分析、识别命名实体、识别特定类型的关系等一系列信息抽取任务。在MUC-7之后,MUC被由NIST引导的ACE(Automatic Content Extraction)[7]评测所取代。

美国国家标准技术研究院(NIST)组织的ACE评测从1999年开始继续进行信息抽取方面的评测。ACE评测1999年7月开始酝酿,2000年12月正式开始启动,迄今已经举办过8次评测,最近正在进行第9次评测(2008年5月)。ACE08[8]所提出的任务包括:单文档内实体探测和识别以及关系探测和识别;跨文档实体探测和识别以及关系探测和识别。文档语种有英语和阿拉伯语两种。在MUC和ACE的促进下,关系抽取技术取得了较大进步,研究热点已从最初的语言学单纯模型的应用发展到使用浅解析器或完全解析器的NLP技术的应用和复杂机器学习方法的应用,而关系抽取性能也有了大幅提升。

国内信息抽取方面的研究虽然起步较晚,但目前也已经在关系抽取方面做出了一些卓有成效的工作。邓擘等人[9]在使用模式匹配技术的基础上引入了词汇语义匹配技术对汉语实体关系进行抽取,并比较了一般模式匹配技术和词汇语义模式匹配技术在汉语实体关系提取任务中的性能。他们的实验结果表明,一般模式匹配技术在处理中文时效果较差,而词汇语义模式匹配技术更适合于处理汉语实体关系抽取任务。姜吉发等人[10]提出了一种自举的二元关系和二元关系模式获取方法BRPAM,该方法能够根据用户初始给出的几个种子二元关系从一个大的自由文本集合中抽取出更多的二元关系。顾雪峰[11]针对文本特征粒度对实体关系识别结果影响较大这一问题,应用动态粒度思想,对识别特征进行逐步细化,构建了一个具有偏序

关系的特征族来进行关系抽取,取得了较好的效果。刘克彬等人[12]实现了基于核函数的中文实体关系自动抽取系统,应用改进的语义序列核函数,结合 KNN 机器学习算法构造分类器来分类并标注关系的类型。通过对 ACE 评测定义的 3 大类 6 子类实体关系的抽取,关系抽取的平均精度达到了 88%。车万翔等人[13]以 2004 年 ACE 评测训练数据作为实验数据,使用两种基于特征向量的机器学习算法 Winnow 和 SVM 进行实体关系抽取,并指出在关系抽取时,应当集中尽力寻找好的特征。Intel 中国研究中心的 ZHANG Yi - Min 和 ZHOU Joe F 等人[14]在 ACL - 2000 上演示了他们开发的一个抽取中文命名实体以及这些实体间相互关系的信息抽取系统,该系统利用基于记忆的学习(Memory - Based Learning,MBL)算法获取规则用以抽取命名实体及它们之间的关系。

3 关系抽取面临的困难

成功的关系抽取取决于正确探测实体,正确判断实体类型,以及正确判断实体间关系的类型。目前对于命名实体的探测以及实体类型的判断技术已经相对成熟,其准确率和召回率一般都能达到 90% 以上。因此能否正确判断实体间关系的类型成为影响关系抽取最终性能的决定性因素。一个比较完整的关系抽取系统应包括依次相连的 5 个模块:NLP 处理和实体抽取、模式匹配或分类、共指消解、新关系处理以及规范化输出。在这个抽取过程中面临的困难基本上可以归纳为以下 3 个方面:

(1)特定领域标引数据集的获取。关系抽取核心部分多采用基于模式匹配和基于机器学习的算法来判断关系是否存在以及关系的类型,而使用这些算法的先决条件是需要预先通过对一个特定领域手工标引的数据集进行学习以获取领域内关系类型的各项特征。此外,这些用于学习的数据集的大小和标引质量都会影响到关系抽取的效果。已标引数据集通常是通过手工标引获得的,这使得特定领域标引数据集的获取比较困难。针对这个问题,一些学者已经提出在关系抽取过程中尽量引入领域 Ontology、领域词表以及 WordNet 等资源来减少关系抽取对已标引数据集的依赖性。此外,学习算法的不断改进也能够减少学习过程中所需的已标引数据量。

(2)模式的获取。基于模式匹配原理的关系抽取方法在很多关系抽取系统中得到了应用。然而,定制特定领域的恰当的关系模式存在较大困难。在以手工方式编制模式过程中,用户必须首先确定给定文集中所有的目标信息表达方式,然后考虑所有的那些表达方式中的变量,最后写出恰当的规则模式。为了使模式编制过程能够更加方便,有学者提出了基于宏的模式编写方式,即给定若干具有特定变量集的宏,当编写领域相关模式时,只要按需要设置宏中的某些变量,就可以自动生成大量相关模式。另外,借助于先进算法的自动模式获取方法也被应用到很多系统中。

(3)共指消解。一个命名实体在文本中可能出现多次,其表现形式也可能不同(例如代名词、反身代词、名词性时间表述等),因此实体间的关系经常被重复探测到。这些指向相同实体间关系的关系实例需要进行合并。在目前的关系抽取系统中,一般使用首语重复法(Anaphors)来解决共指消解问题。通过首语重复法,相关联的实体被合并,并在候选短语表中选择一个最恰当的表达形式。

4 关系抽取的几种技术方法

针对关系抽取过程中的难题,信息抽取领域的学者们进行了长期探索和不懈努力。到目前为止,已经有许多关系抽取方法被应用在各种实验系统当中。这些方法所遵循的技术方法基本可以归纳为:基于模式匹配的关系抽取、基于词典驱动的关系抽取、基于机器学习的关系抽取、基于 Ontology 的关系抽取以及混合抽取方法。

4.1 基于模式匹配的关系抽取

在关系抽取研究领域,普遍使用基于模式匹配的关系抽取方法。这种抽取方法通过运用语言学知识,在执行抽取任务之前,构造出若干基于语词、基于词性或基于语义的模式集合并存储起来。当进行关系抽取时,将经过预处理的语句片段与模式集合中的模式进行匹配。一旦匹配成功,就可以认为该语句片段具有对应模式的关系属性。

在应用基于模式匹配的关系抽取方法时,最困难的步骤是关系模式的建立。最初关系模式的建立需要依靠语言学家对抽取任务涉及的领域语料进行深入分析,借鉴已有语言学成果,穷举各种可能的关系表达,

手工编制关系模式。这样的方法一方面使编制模式的周期太长,应用成本很高;另一方面,当抽取系统被用来进行新领域的关系抽取时,就需要语言学家根据新的领域抽取特点重新编制关系模式,这在现实应用中实现起来非常困难。针对这一问题,一些学者提出了不同的解决思路。

Douglas E. Appelt 等人[15]在 MUC-6 上提出的 FASTUS 抽取系统中,通过引入"宏"的概念将各种领域依赖规则以一种具有扩展性的、通用方式表达。用户只需要修改相应"宏"中的参数设置,就可以快速配置好特定领域任务的关系模式规则。FASTUS 系统中的所有模式规则被分成领域依赖和领域独立两部分。领域独立部分可以看作确定参数的宏。这些模式规则在一个相对粗的粒度层次上覆盖各种句法结果,目标是要对于符合模式的动词构造恰当的谓词-参数(Predicate-argument)关系。领域依赖的规则包含一些参数,这些参数必须通过"宏"的实例化来产生实际模式规则。这些领域依赖规则会指定哪一个动词载有领域相关信息,以及这些参数的领域依赖限制以及规则的语义。FASTUS 系统采用的编译时转换的方式实现了使用 12 个宏规则和 15 个领域依赖的规则就可以实现大概 100 个明确表达的模式的效果,这为系统在处理领域关系抽取任务时的配置工作节约了大量时间。

Roman Yangarber 等人[16]在 MUC-7 上提出的 Proteus 抽取系统采用了基于样本泛化的关系抽取模式构建方法。用户通过 Proteus 系统提供的模式构建界面,对含有某种关系的例句进行分析,识别出所含关系的要素,并将这些要素泛化,最后经用户确认存储经泛化表达的模式。系统还会应用集成的 Meta-rules,从用户生成的简单的主动句模式或独立的名词短语产生一组句法转换器,例如某词的被动词、关系词,以及被动关系、减少关系模式等。Proteus 也能将可选修饰部分插入到产生的变量(例如:临近句子等),来扩展模式的覆盖范围。

4.2 基于词典驱动的关系抽取

与基于模式匹配的关系抽取方法相比,基于词典驱动的关系抽取方法显得非常灵活。新的关系类型能够仅仅通过向词典添加对应的动词入口而被抽取。用户不需要具备复杂的模式语言知识就可以轻松配置抽取系统。

Chinatsu Aone 等人[17]在 MUC-7 上提出了一个快速、灵巧的大规模事件和关系抽取系统(Large-Scale Relation and Event Extraction System,REES)。该系统采用的基于词典驱动的关系抽取方法旨在能够抽取尽可能多类型的关系和事件,但耗费的努力最小,准确率较高。在 REES 系统中,当输入语料经过名称标识和名词短语标识阶段的处理,形成基于 XML 的输出。接着关系识别模块应用词典驱动模型,通过基于句法的一般模式来识别关系和事件。REES 的词典驱动方法需要对于每一个事件指示词设置一个词典入口,而这个词通常是动词。词典入口具体化了该动词参数的句法和语义限制。

基于词典驱动的关系抽取方法的缺点也非常明显。它只能识别以动词为中心词的关系,而对于名词同位语之类的关系抽取就很难实现了。另外,使用这种方法无法对系统中没有对应词汇入口的新关系进行探测。

4.3 基于机器学习的关系抽取

基于机器学习的关系抽取方法是目前应用比较广泛的方法。该方法实质是将关系抽取看作是一个分类问题。通过具体的学习算法,在人工标引语料的基础上构造分类器,然后将其应用在领域语料关系的类别判断过程中。目前使用比较多的学习算法有 MBL 算法和 SVM 算法。

Intel 中国研究中心的 ZHANG Yi-Min 和 ZHOU Joe F 等人[18]在 ACL-2000 上演示了他们开发的一个抽取中文命名实体以及这些实体间相互关系的信息抽取系统,该系统就是利用 MBL 算法获取规则用以抽取命名实体及它们之间的关系。ZHANG Yi-Min 等人将中文实体名和关系识别看作一系列分类问题。整个过程能够被分成两个阶段:第一阶段是学习过程,若干分类器从训练数据构建起来;第二阶段是抽取过程,通过使用学习得到的分类器抽取中文实体名和它们的关系。之所以选择 MBL 作为学习算法,是因为它非常适合处理从大量不同来源获取的特征,并且能记住例外案例和低频案例,而这对于后续的推断阶段非常有用。该系统已经能够抽取的关系类型包括 Employee-of,Location-of,Product-of,和 No-relation。通过提供更多的训练数据,能够轻易扩展关系抽取类型。

Zhu Zhang[19]提出的基于SVM的弱监督关系分类系统应用SVM算法进行关系抽取。Zhu Zhang提出的弱监督学习过程包括两个组件:一个底层监督学习器和一个在其上的Bootstrapping算法。底层监督学习器是一个支持向量分类器,它使用从当前可获得的已标注数据训练而来的模型,对未标记的数据进行分类。Bootstrapping算法则负责选择最有可能被正确标记的实例,并通过使用它们来增强已标记数据的训练效果。该系统在进行分类任务时用到了词语特征、浅句法特征、深层句法特征以及序列标志、实体类型等特征。

Michele Banko等人[20]提出了一个新颖的开放信息抽取方法(Open IE,OIE),实现了对网络上海量异构信息中可能存在的关系的抽取。该方法既不需要手工标注训练集作为训练语料,也不局限于特定领域,而是通过自动学习和统计来实现关系抽取。开放信息抽取方法的实现分为3个阶段:通过对一个相对较小的语料集进行深层解析,自动抽取并标注可信的和不可信的关系三元组。这些三元组的特征向量被作为训练样例进行幼稚贝叶斯分类器的训练;在训练好的分类器上进行大量网络文献的关系抽取。为了确保较高的处理效率,抽取器并不使用解析器对文献进行深层解析,而是将较容易获得的词性标注、序列等特征作为分类器的输入。这一阶段的输出是去除了不必要的修饰词后的候选关系三元组集合;对这些候选三元组进行合并,通过统计的方法计算各个关系三元组的可信度,并建立索引。

4.4 基于Ontology的关系抽取

知识管理过程中,利用信息抽取技术抽取的实体以及实体间的关系来构建和丰富本体,是一种行之有效的方法。另一方面,借助已有的本体层次结构和其所描述的概念之间的关系来协助进行关系的抽取,也不失为一种行之有效的关系抽取方法。

José Iria等人[21,22]提出了一个基于本体的关系抽取通用软件框架—可训练关系抽取框架(Trainable Relation Extraction Framework,T-Rex)。设计该框架的目的是要提供语义网自动化语义标注任务需要的灵活度。由于T-Rex采用了参数化的插件结构,因此可以对多种基于不同抽取算法的插件进行集成和测试。T-Rex最具特色的地方是它采用了规范的基于图的数据模型。该数据模型借助本体实现等级层次的表达结构,并允许以一致的方式任意链接子图,例如共指关系链接,语法关系链接,与HTML格式相关的链接等。T-Rex数据模型的表示是等级化的,能够将语料模型化到字符级、语词级、短语级、语句级和文档级层次。通过对本体的定义和扩充,可以实现使用该多层次数据模型对于语料的多种特征集表达的一致性。

Alexander Schutz等人[1]将DOLCE、SUMO、SportEventOntology等本体有机结合在一起来描述足球领域的相关概念及概念之间的关系,建立了能够自动识别高相关三元组(概念对和概念之间的关系)的RelExt系统。该系统通过从文本集合抽取相关词项和动词,借助语言学和统计学处理过程计算词项之间的相关关系。该系统目前能够处理1 570个足球领域相关概念(类)和487个直接关系。

Marta Sabou和Mathieu d'Aquin等人[23]提出的SCARLET系统通过自动选择和查询本体的方法来发现概念实体之间的关系。例如,当要确定两个概念实体Researcher和AcademicStaff之间的关系时,SCARLET先识别网络上能够提供上述概念实体相关信息的本体,然后综合这些信息来推断概念实体之间的关系。当上述两个概念实体已经在某个本体中被定义,这两个实体间的关系就可以通过本体获取到。如果上述的概念实体在不同的本体中被描述,例如在一个本体中描述了Researcher属于ResearchStaff的关系,在另一个本体中描述了ResearchStaff属于AcademicStaff的关系,则通过关系逻辑推导可以获取Researcher和AcademicStaff之间的关系。

4.5 混合抽取方法

在关系抽取研究的初期阶段,无论是基于词典的抽取方法还是基于模式的抽取方法,都仅将一种抽取方法作为整个关系抽取过程的核心。随着关系抽取研究的不断深入,研究者逐渐意识到,单纯的抽取方法在识别特征和识别模式方面难以避免地会具有局限性。为了将更多的已有关系识别特征加入到关系抽取过程中来,一些将多种现有关系抽取方法相结合的混合抽取方法被提出来。其中具有代表性的是Lucia Specia和Enrico Motta[24]提出的一个抽取语义关系的混合方法。

该方法通过管道(Pipeline)方式引入了解析器(Parser),词性标注器(Part-of-speech Tagger),命名实体识别系统,基于模式的分类器以及词义辨析模块,

并用到了领域本体,知识库以及词语数据库等资源。

该方法的核心策略是匹配一个语言学三元组和他们对应的语义组件。这不仅包括匹配关系,还包括匹配这些关系相关联的项。语言学三元组的探测包括一系列的语言学处理步骤。词项和概念的匹配通过一个领域本体和一个命名实体识别系统引导。关系识别依赖于在领域本体和词库中的知识,以及基于模式的分类和词义辨析模块。除了抽取已经在领域本体中存在的关系,该框架还可以通过模式匹配策略来发现词项类型之间的新关系。

5 结 语

经过20多年的发展,关系抽取理论和方法愈加完善。从最初的手工编写模式和词典进行关系抽取,发展到目前借助Ontology和知识库等多种知识资源的综合关系抽取,关系抽取的正确率和召回率在不断提高,对不同领域的适应性也在不断加强。目前仍然存在一些比较实际的问题阻碍了关系抽取在实际中的应用,这包括已标引数据集的获取、关系模式的构建、共指消解等问题。随着这些问题的进一步解决,关系抽取技术必然会在增强检索系统功能、语义网标注、本体学习等领域得到广泛应用。

参考文献:

[1] Schutz A, Buitelaar P. RelExt:A Tool for Relation Extraction from Text in Ontology Extension[C]. *4th International Semantic Web Conference*, Galway, Ireland, November 6-10, 2005:593-606.

[2] Katrenko S, Adriaans P. Learning Relations from Biomedical Corpora Using Dependency Tree Levels[C]. In:*Proc. BENELEARN conference*(2006), 2006.

[3] Relationship Extraction[EB/OL]. [2008-05-30]. http://en.wikipedia.org/wiki/Relationship_extraction.

[4] The ACE 2004 Evaluation Plan[EB/OL]. [2008-05-30]. http://www.nist.gov/speech/tests/ace/2004/doc/ace04-evalplan-v7.pdf.

[5] Automatic Content Extraction 2008 Evaluation Plan (ACE08)[EB/OL]. [2008-05-30]. http://www.nist.gov/speech/tests/ace/2008/doc/ace08-evalplan.v1.2.pdf.

[6] MUC[EB/OL]. [2008-05-30]. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/.

[7] ACE[EB/OL]. [2008-05-30]. http://www.nist.gov/speech/tests/ace/.

[8] ACE08 Annotation Tasks[EB/OL]. [2008-05-30]. http://projects.ldc.upenn.edu/ace/annotation/.

[9] 邓擘,樊孝忠,杨立公. 用语义模式提取实体关系的方法[J]. *计算机工程*,2007,33(10):212-214.

[10] 姜吉发,王树西. 一种自举的二元关系和二元关系模式获取方法[J]. *中文信息学报*,2005,19(2):71-77.

[11] 顾雪峰. 基于动态粒度思想的实体关系识别方法研究[EB/OL]. [2008-05-30]. http://www.cnki.com.cn/grid20/Detail.aspx.

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

[13] 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.

[14] Zhang Y M, Zhou J F. A Trainable Method for Extracting Chinese Entity Names and Their Relations[C]. In: *Proceedings of the Second Chinese Language Processing Workshop*, Hong Kong, 2000:66-72.

[15] Appelt D E, Hobbs J R, Bear J, et al. SRI International FASTUS System:MUC-6 Test Results and Analysis[C]. In: *Proceedings of the 6th Message Understanding Conference* (*MUC-6*), 1995:237-248.

[16] Roman Y, Grishman R. NYU:Description of the Proteus/PET System as Used for MUC-7 ST[C]. In: *Proceedings of the 6th Message Understanding Conference* (*MUC-7*), 1998.

[17] Aone C, Ramos2Santacruz M. Rees:A large-scale relation and event extraction system[C]. In:*Proc of the 6th Applied Natural Language Processing Conference*, New York, 2000:76-83.

[18] Zhang Y, Zhou J F. A Trainable Method for Extracting Chinese Entity Names and Their Relations[C]. In: *Proceedings of the second Chinese Language Processing Workshop*, ACL, 2000:66-72.

[19] Zhu Z. Weakly-supervised Relation Classification for Information Extraction[C]. In:*Proceedings of the Thirteenth ACM conference on Information and Knowledge Management*, Washington D. C., 2004:581-588.

[20] Banko M, Cafarella M J, Soderland S, et al. Open Information Extraction from the Web[C]. In: *Proceeding of the International Joint Conferences on Artificial Intelligence*, 2007.

[21] Iria J. T-Rex:A Flexible Relation Extraction Framework[C]. In: *Proceeding of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics* (*CLUK'05*), Manchester, January 2005.

[22] Iria, Mr. José, Ciravegna, Fabio. Relation Extraction for Mining the Semantic Web[C]. In: *Proceedings Machine Learning for the Semantic Web Dagstuhl Seminar* 05071, Dagstuhl, 2005.

[23] Sabou M, Mathieu d'Aquin, Motta E. SCARLET:SemantiC relAtion DiscoveRy by Harvesting onLinE onTologies[C]. In: *Proceedings of the 5th European Semantic Web Conference*, June, 2008.

[24] Specia L, Motta E. A Hybrid Approach for Extracting Semantic Relations from Texts[EB/OL]. [2008-05-30]. http://www.dcs.shef.ac.uk/~lucia/publications/SpeciaMotta_OLP2-2006.pdf.

(作者E-mail:xujian@mail.las.ac.cn)

实体关系抽取的技术方法综述

作者： 徐健， 张智雄， 吴振新， Xu Jian， Zhang Zhixiong， Wu Zhenxin

作者单位： 徐健,Xu Jian(中国科学院国家科学图书馆,北京,100190;中国科学院研究生院,北京,100049;中山大学资讯管理系,广州,510275)， 张智雄,吴振新,Zhang Zhixiong,Wu Zhenxin(中国科学院国家科学图书馆,北京,100190)

刊名： 现代图书情报技术 PKU CSSCI

英文刊名： NEW TECHNOLOGY OF LIBRARY AND INFORMATION SERVICE

年，卷(期)： 2008，""(8)

被引用次数： 0次

参考文献(24条)

1. Schutz A. Buitelaar P RelExt:A Tool for Relation Extraction from Text in Ontology Extension 2005

2. Katrenko S. Adriaans P Learning Relations from Biomedical Corpora Using Dependency Tree Levels 2006

3. Relationship Extraction 2008

4. The ACE 2004 Evaluation Plan 2008

5. Automatic Content Extraction 2008 Evaluation Plan(ACE08) 2008

6. 查看详情 2008

7. 查看详情 2008

8. ACE08 Annotation Tasks 2008

9. 邓擘. 樊孝忠. 杨立公 用语义模式提取实体关系的方法[期刊论文]-计算机工程 2007(10)

10. 姜吉发. 王树西 一种自举的二元关系和二元关系模式获取方法[期刊论文]-中文信息学报 2005(02)

11. 顾雪峰 基于动态粒度思想的实体关系识别方法研究 2008

12. 刘克彬. 李芳. 刘磊 基于核函数中文关系自动抽取系统的实现[期刊论文]-计算机研究与发展 2007(08)

13. 车万翔. 刘挺. 李生 实体关系自动抽取[期刊论文]-中文信息学报 2005(02)

14. Zhang Y M. Zhou J F A Trainable Method for Extracting Chinese Entity Names and Their Relations 2000

15. Appeh D E. Hobbs J R. Bear J SRI International FASTUS System:MUC-6 Test Results and Analysis 1995

16. Roman Y. Grishman R NYU:Description of the Proteus/PET System as Used for MUC-7 ST 1998

17. Aone C. Ramos2Santacruz M Rees:A large-scale relation and event extraction system 2000

18. Zhang Y. Zhou J F A Trainable Method for Extracting Chinese Entity Names and Their Relations 2000

19. Zhu Z Weakly-supervised Relation Classification for Information Extraction 2004

20. Banko M. Cafarella M J. Soderland S Open Information Extraction from the Web 2007

21. Iria J T-Rex:A Flexible Relation Extraction Framework 2005

22. Iria,Mr. José. Ciravegna,Fabio Relation Extraction for Mining the Semantic Web 2005

23. Sabou M. Mathieu d'Aquin. Motta E SCARLET:SemantiC relAtion DiscoveRy by Harvesting onLinE onTologies 2008

24. Specia L. Motta E A Hybrid Approach for Extracting Semantic Relations from Texts 2008

相似文献(10条)

1. 会议论文 张素香. 李蕾. 钟义信 基于自由文本的中文实体关系抽取研究 2005

针对信息抽取技术发展状况,本文基于信息抽取技术的发展历史,总结了目前世界上主要的信息抽取系统主要使用的相关技术和方法,分析它们的优缺点.在此基础上,结合全信息理论和机器学习,提出了基于全信息中文实体关系抽取模型,并对各个模块进行了详细地分析和阐述.

2.学位论文 刘克彬 基于核函数的命名实体关系抽取技术研究 2006

随着计算机的普及以及互联网的迅猛发展，大量的信息以电子文本的形式出现在人们面前。为了应对信息爆炸带来的挑战，迫切需要一些自动化的工具帮助人们在海量信息源中迅速找到真正需要的信息。信息抽取(Information Extraction)研究正是在这种背景下产生的。信息抽取的主要目的是将无结构的文本转化为结构化或半结构化的信息，并以数据库的形式存储，供用户查询以及进一步分析利用。

信息抽取有三个基本任务，命名实体识别、实体关系抽取和事件发现。实体关系抽取不仅是信息抽取的一项重要任务，也是事件发现和多种应用系统的基础，具有重要意义。实体关系抽取的基本任务是寻找并判定实体对之间存在的特定关系。当前主要的抽取技术可分为基于知识库的抽取算法、基于特征向量的机器学习算法、基于核函数的机器学习算法、基于模式的Bootstrapping算法。

本文的工作在命名实体识别的基础上重点研究了实体关系抽取技术并实现了一个完整的实体关系抽取系统。通过深入分析关系抽取技术的重点和难点以及现有技术的特点和不足，设计实现了基于改进的语义核函数的关系抽取系统。

本文的成果和贡献主要体现在以下几个方面：

1)命名实体识别算法：作为关系抽取的前续工作，命名实体识别是本文工作的一个重要组成部分。本文的命名实体识别算法采用字典结合训练规则的方式，具有很高的准确率和召回率。

2)基于核函数的关系抽取算法的研究和改进。这部分是本文的主要工作，包括几个部分：

A)首先是对现有的核函数进行归类，研究它们的优点和不足并加以总结。

B)选择具有多种优良性质的序列核函数作为主要研究对象，对其进行了较大的改进，得到一种语义序列核函数。这里的语义核函数指的是将语义知识嵌入到核函数的计算过程中，在不增加计算复杂度的情况下显著提升了学习算法的分类以及泛化能力。

C)语义知识的获取也是比较重要的一部分工作，本文的语义知识获取充分利用了著名的中文语义本体Hownet。

D)实现了基于语义核函数的KNN学习算法并应用于关系抽取系统，与其他关系抽取系统相比，本文的方法具有较高的准确率以及良好的泛化能力。

E)扩展现有的二元实体关系为三元关系，使关系抽取结果包含更为丰富的信息。

3)命名实体关系抽取系统实现：本文介绍的关系抽取系统采用模块化设计，总共包括8个主要的功能模块。这些模块都具备两种不同的实现方式，一种是Gate自然语言处理平台下插件形式的实现，一种是独立的Java应用实现。因此本文的整个系统既可以作为Java独立应用运行，也可以作为Gate环境下的插件自由组合以满足各种不同应用的需要。

3.学位论文 张婷 基于迭代方法的命名实体关系抽取技术研究 2008

随着互联网的普及,信息的数量与日俱增,人们需要从海量的信息中提取真正需要的信息,信息抽取的研究正是在这种背景下产生的。信息抽取的主要目的是将无结构的文本转化为结构化或半结构化的信息,并以数据库的形式存储,供用户查询以及进一步分析利用。

信息抽取的基本任务包括命名实体识别和实体关系抽取。其中命名实体识别是实体关系抽取的基础,实体关系抽取是事件发现和多种应用系统的基础。实体关系抽取的任务是寻找并判定实体对之间存在的特定关系。当前主要的抽取技术可分为基于知识库的抽取算法、基于特征向量的机器学习算法、基于核函数的机器学习算法、基于模式的Bootstrapping算法。

本文的工作在命名实体识别的基础上重点研究了实体关系抽取技术并实现了一个实验性的实体关系抽取系统。通过深入分析关系抽取技术的重点和难点以及现有技术的特点和不足,设计实现了特定关系抽取检索查询系统。本文的主要工作体现在以下几个方面：

1)命名实体识别算法：作为关系抽取的前续工作,命名实体识别是本文工作的一个重要组成部分。本文的命名实体识别算法重点针对机构实体的识别,采用字典结合规则的方式,其中利用互信息原理对机构名简称进行识别,具有很高的准确率。

2)特定关系抽取及三元命名实体扩展算法：根据规则提取了某类特定实体关系,并对提取到的实体采用规则加迭代方式进行了关系扩展,具有很好的效果。

3)命名实体特定关系检索查询系统实现：本文介绍的命名实体特定关系检索查询系统采用模块化设计,总共包括六个主要的功能模块。

4.期刊论文 牟晋娟.包宏.MU Jinjuan.BAO Hong 中文实体关系抽取研究 -计算机工程与设计2009,30(15)

针对基于特征向量的实体关系抽取方法中特征向量一般构造方法存在的不足,提出了基于互信息的实体对特征向量构造方法.该方法引入词和实体关系类别之间的互信息作为一个句子中实体对左右两边上下文特征提取的判断标准,并对实体关系类别特征词条进行编码,在此基础上再对实体对左右两边的上下文信息进行编码.这样做压缩了实体对上下文信息编码的维数,突出了实体关系各类别特性.实验结果表明本文的实体关系特征向量构造方法提高了中文实体关系抽取的准确率和召回率.

5.学位论文 聂昆 网络舆情资源的信息抽取研究 2009

随着技术的快速发展，互联网已经成为民众表达意愿的主要平台。网络舆情是公众对自己关心或与自身利益紧密相关的各种公共事务所持有的多种情绪、态度和意见交错的总和，它不仅是民意在互联网上的再现，而且对现实社会也具有越来越大的影响力，近年来网络舆情受到的关注越来越多，已经成为一个多学科的交叉研究领域。

本文从各学科网络舆情研究现状总结入手，站在情报学角度对网络舆情资源的现状进行分析，重新梳理了资源的分类体系，并提出了面向网络舆情资源的信息抽取方法。本文的主要工作有：

1. 总结当前网络舆情的研究状况，从理论基础和技术手段两方面总结了各学科对该领域的贡献，提出情报学可以深入研究的问题。

2. 从情报学研究的角度出发，结合技术手段对网络舆情资源重新梳理。总结了目前网络舆情资源的分类、特点，并对获取资源的方法和难度进行了探讨。在资源分析上，对社会学提出的理论进行了补充和调整。

3. 在总结了资源特点的基础上，提出了面向主题的自由文本形式的网络舆情信息抽取方法，并进行了实验。该方法独立于资源的形式姿态，抽取出的实体关系结果可以揭示舆情信息，并为信息检索等后续工作服务。

6.会议论文 徐芬.王挺.陈火旺 基于SVM方法的中文实体关系抽取 2007

实体关系抽取是很多自然语言处理任务的重要基础。本文针对中文中实体关系的特点,设计了一系列的特征。包括词、词性标注、实体和出现信息、包含关系和知网提供的概念信息等,以构成实体间关系的上下文特征向量并使用SVM方法进行了中文实体关系抽取。以ACE2004的训练语料作为实验数据,得到了较好的识别性能。同时根据分级实验的结果,考察了各种特征集对识别性能的影响,得到下一步研究的方向。

7.学位论文 彭学政 基于统计方法的中文命名实体识别与关系抽取 2008

随着计算机的广泛应用和互联网技术的迅猛发展，社会的信息总量呈指数级增长。面对信息爆炸带来的挑战，亟需一些智能化的工具来帮助用户获取真正有用的信息，信息抽取正是在这种背景下提出的，并已经成为当前研究的一个热点问题。信息抽取的主要目的是将无结构或者半结构化的文本转化为结构化的信息，其研究任务可分为：命名实体识别、实体关系抽取、指代消解和事件探测这四个主要研究点。本文针对中文信息抽取当中的命名实体识别和实体关系抽取技术展开研究，主要的贡献有如下几点：

1、提出了一种外部词典与统计相结合的汉语分词方法。该方法利用外部词典来改进“由字构词”的汉语分词方法，既保留了传统词典分词方法对词典词的处理精度高的优点，又具有统计方法汉语分词方法在未登录词处理上的优势，有效地提升了对词典词的处理能力，从而地提高了汉语分词的整体性能。同时，该方法只需要在较小的标注语料库上训练，就能获得令人满意的分词结果，从而减轻了统计方法对于标注语料库的依赖性。

2、尝试了专家知识与机器学习相结合的中文命名实体识别方法。针对现有命名实体识别方法的不足，本文采用了比最大熵马尔可夫模型(MEMM)更加优越的条件随机场模型(CRFs)作为机器学习的主要框架，通过利用人名、地名和机构名的构成规则等专家知识，以及人名姓氏和名字常用字列表、地名常用后缀列表、机构名常用后缀列表等词典资源来辅助机器学习，提高了命名实体识别的准确率和召回率。

3、研究了一种面向主题的实体关系抽取方法。提出利用依存句法分析的结果树来计算两个命名实体之间的“语法距离”，削弱了汉语当中修饰语在计算实体之间的距离时的干扰作用，提高了关系抽取的准确率。同时，本文还结合词性、词在文字窗口中的位置、词之间的依存关系等信息来进行关系描述词语的抽取，从而能够自动为所抽取的实体关系对赋予较为准确的标签。

4、在以上研究成果的基础上，设计并开发了一个实用的关系抽取模块，应用于互联网舆情监测系统，取得了很好的实际应用效果。

8.期刊论文 周峰.吴斌.石川 复杂网络构建中信息抽取技术综述 -数字图书馆论坛2008,""(6)

复杂网络为我们研究复杂性问题提供了一个新的视角和方法,激起了对于不同的实际网络特性的研究热潮.同时,信息抽取作为一门逐渐成熟的技术,在信息处理自动化中具有基础性的地位.将信息抽取和复杂网络研究相融合,通过信息抽取技术,可以抽取到节点信息、边的信息,为复杂网络的构建提供基本的数据准备,大大扩展了复杂网络的应用.文章首先介绍了信息抽取的基本概念和类型等,随后对复杂网络构建中主要的信息抽取技术作了简单的描述和分析.

9.学位论文 李晶 基于网络抱团发现的命名实体关系抽取 2006

关系抽取是是信息抽取研究领域的一个重要课题。关系抽取的目的是从文本中发现两命名实体间的关系。近年来，该技术得到越来越多的关注，被运用到各个领域中，如：信息抽取，本体构造，问答系统，生物技术等。

自从1995年，第六届信息理解会议(theSixthMessageUnderstandingConferenceMUC-6)提出关系抽取这个概念以来，在命名实体关系抽取方面已经开展了大量的研究工作，但绝大部分研究都是基于有导学习的。有导学习方法最大的问题在于需要花大量的时间去标注足够数量的训练语料，此外系统很难从一个领域移植到另外一个领域中。为了突破有导学习方法中的限制，无导方法被提出来。但是，目前无导方法仍然存在一些问题：(1)命名实体对一般是通过一定范围内的上下文来进行描述地，但是如何设置上下文窗口大小，一直以来都没有一个客观的标准；(2)在命名实体对的聚类过程中，往往存在着很大的噪音，如何能在噪音存在的情况下保证较好的聚类效果；(3)命名实体对之间的语义关系是具有层次结构的，如何描述这种层次结构的关系。

针对以上问题，本文尝试性地提出了一种基于网络化数据挖掘的命名实体对关系抽取的方法。在该方法中，我们采用了如下三种关键技术：1、利用网络化结构来表示命名实体对；2、基于抱团现象的命名实体对聚类；3、基于语义层次的命名实体关系描述。特别值得一提的是在关键技术1中我们着重解决了实体对上下文窗口大小的设置的问题，在关键技术2中我们创新性的提出了一个如何在带权网络中发现抱团现象的方案。

为了验证提出的命名实体对关系抽取方法的可行性和有效性，我们以半年人民日报为语料进行实验。结果表明，本文提供的方法不但可以以较高的准确率发现命名实体间的语义关系，而且能够恰当地标注命名实体间的语义关系。

10.会议论文 车万翔.刘挺.李生 实体关系自动抽取 2004

实体关系抽取是信息抽取研究领域中的重要研究课题.本文使用两种基于特征向量的机器学习算法,Winnow和支持向量机(SVM),在2004年ACE(Automatic Content Extraction)评测的训练数据上进行实体关系抽取实验.两种算法都进行适当的特征选择,当选择每个实体的左右两个词为特征时,达到最好的抽取效果,Winnow和SVM算法的加权平均F-Score分别为73.08％和73.27％.可见在使用相同的特征向量,不同的学习算法进行实体关系的识别时,最终性能差别不大.因此使用自动的方法进行实体关系抽取时,应当集中精力寻找好的特征.