关系抽取技术研究综述

黄 勋 游宏梁 于 洋
(中国国防科技信息中心 北京 100142)

【摘要】对关系抽取技术研究概况进行总结。在回顾关系抽取发展历史的基础上,将关系抽取研究划分为两个阶段:面向特定领域的关系抽取研究和面向开放互联网文本的关系抽取研究。在分析相关文献的基础上,总结出两个研究阶段的技术路线:面向特定领域的关系抽取技术以基于标注语料的机器学习方法为主;面向开放互联网文本的关系抽取则根据不同任务需要,采取基于启发式规则的方法或者基于背景知识库实例的机器学习方法。

【关键词】关系抽取 信息抽取 机器学习
【分类号】TP391

A Review of Relation Extraction

Huang Xun You Hongliang Yu Yang
(China Defense Science & Technology Information Center, Beijing 100142, China)

【Abstract】The paper summarizes the research of relation extraction techonology. It firstly gives a brief overview of relation extraction, and divides the research into two phases: the relation extraction in specific domains and the relation extraction of Web text. Then, analyzes the major methodologies of the two phases: the relation extraction in specific domains mainly uses machine learning methods with annotated corpora, while the relation extraction of Web text uses rule - based methods or distant supervision methods according to different demands.

【Keywords】Relation extraction Information extraction Machine learning

1 引 言

随着信息技术的快速发展和计算机的普及,各种各样的信息在互联网上迅猛增加。在信息爆炸的时代,如何从海量信息中快速准确地获取用户感兴趣的信息已经成为亟待解决的问题。在这种背景下,信息抽取(Information Extraction, IE)技术应运而生。关系抽取(Relation Extraction, RE)是信息抽取的关键技术之一,近年来逐渐受到研究界的广泛关注。

信息抽取包含三个关键技术:实体抽取、关系抽取、事件抽取。其中实体抽取是关系抽取和事件抽取的基础,旨在从文本中识别出人名、地名、机构名、日期、数额等实体信息。为了深入理解自然语言文本信息,要在实体识别的基础上,抽取出这些实体之间存在的语义关系。这项抽取实体间语义关系的任务,即关系抽取。实体间的关系可被形式化描述为关系三元组 $<Entity_1, Relation, Entity_2>$,其中 $Entity_1$ 和 $Entity_2$ 是实体类型,Relation 是关系描述。关系抽取即从自然语言文本中抽取出关系三元组 $<Entity_1, Relation, Entity_2>$,从而提取文本信息。

关系抽取技术在海量信息处理、知识库自动构建和搜索引擎等领域具有重要意义:通过关系抽取技术,从无结构的自然语言文本中抽取出格式统一的关系数据,有助于计算机快速处理大规模文本数据,提高处理效率;通过抽取实体之间的语义关系,能够建立多个实体之间广泛的信息关联,有助于建立领域本体,促进知识图谱的构

收稿日期:2013 - 07 - 12
收修改稿日期:2013 - 10 - 25

建;通过深入挖掘和分析自然语言文本中的语义关系信息,能够进一步理解和匹配用户的查询意图,从而为用户提供更精准的搜索服务。由此可见,关系抽取技术不仅仅具有深刻的理论意义,而且具有广阔的应用前景。

2　关系抽取的历史和评价体系

2.1　关系抽取的历史

20 世纪 80 年代末以来,美国国防高级研究计划局(Defense Advanced Research Projects Agency, DARPA)主持召开了消息理解会议(Message Understanding Conference, MUC)[1],通过测评驱动的会议模式推动了信息抽取研究的蓬勃发展。MUC 会议一共举办了 7 届,定义了包括命名实体识别(Named Entity Recognition, NER)、模板关系(Template Relation, TR)、情节模板(Scenario Template, ST)等文本挖掘任务。其中关系抽取任务在 1998 年 MUC - 7[2] 会议上被引入,目的是通过填充关系模板槽的形式抽取出实体之间存在的 Location_of、Employee_of 和 Product_of 三大类关系。MUC 系列测评会议对于关系抽取的研究起到了巨大的推动作用。

自从 MUC 会议于 1998 年停办后,美国国家标准与技术研究院(National Institute of Standards and Technology, NIST)组织开展了自动内容抽取(Automatic Content Extraction, ACE)测评会议[3]。ACE 会议旨在研究自动抽取出新闻语料中的实体、关系以及事件等内容。关系抽取属于 ACE 会议定义的关系检测与识别(Relation Detection and Recognition, RDR)任务[4]。ACE 会议提供了关系抽取的测评语料,也构建了详细的实体关系类型,将关系抽取任务进一步细化。

ACE 会议于 2009 年并入美国国家标准与技术研究院组织的国际文本分析会议(Text Analysis Conference, TAC)后,关系抽取并入知识库构建(Knowledge Base Population, KBP)领域的槽填充(Slot - Filling)任务[5]。TAC KBP 会议推动了面向知识库构建过程的关系抽取研究,很多研究者开始利用大规模的开源知识库,采取基于 Distant Supervision[6] 的方法研究关系抽取。

随着互联网的发展,出现了海量异构而且含有噪声的数据,传统测评会议定义的面向特定领域和特定关系的关系抽取任务已经不再适应新的需求。为了解决互联网海量数据的文本挖掘和分析任务,很多学者开始研究开放式信息抽取(Open Information Extraction, OpenIE)技术[7]。开放式实体关系抽取作为其中的重要子任务和关键技术,受到了研究者的广泛关注。

关系抽取技术发展呈现出两个阶段:

(1)机器学习理论的引入,促使关系抽取由基于语法规则向有监督机器学习方法转变。ACE 等会议提供了标准的关系体系、训练语料和测试样例,促进了有监督机器学习理论在关系抽取中的应用。研究者们普遍将关系抽取视为分类问题,研究重点是语形层面的特征工程。大量研究表明了支持向量机、条件随机场等模型在关系抽取研究中的有效性,同时也发掘了一批适合关系抽取任务的词汇和句法的语形特征。

(2)研究领域向互联网开放领域的拓展,促进了基于半监督和无监督关系抽取技术的发展。知识库构建是关系抽取的重要应用场景,维基百科、YAGO 和 Freebase 等背景知识库所蕴含的广大事实型信息缓解了标注语料不足的问题,因此基于 Distant Supervision 的抽取方法取得了一定效果。但是互联网开放语料的规模更大、包含的关系类型更复杂,直接面向开放语料进行抽取具有更大的实际意义。由于开放语料中绝大部分未经人工标注,因此以半监督和无监督方法为主的开放式关系抽取逐渐成为研究重点。除了利用语形特征之外,开放式关系抽取研究还引入语义特征,从而提升了关系抽取在大规模语料上的适用性。

2.2　关系抽取的评价体系

ACE 会议的 RDC 任务将关系分为 5 到 7 个大类,10 到 20 多个子类。以 ACE 2008[8] RDC 任务为例:关系大类为 ART(Artifact)、GEN - AFF(General Affiliation)、METONYMY、ORG - AFF(Org - Affiliation)、PART - WHOLE(Part - to - Whole)、PER - SOC(Person - Social)和 PHY(Physical)7 类,分别包含 User - Owner - Inventor - Manufacturer、Org - Location、Employment、Subsidiary、Family、Located 等 18 个子类。ACE 测评根据对所有关系类别的平均识别效果判定系统性能。由于 ACE 会议提供了公开标准的测评平台,传统的关系抽取研究一般将关系类别限定在 ACE RDC 任务定义的关系类别上。

针对每个关系类别,一般用准确率(Precision)、召

回率(Recall)和 F_1 值来衡量抽取结果:

$$\text{Precision} = \frac{C}{M} \quad (1)$$

$$\text{Recall} = \frac{C}{N} \quad (2)$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})} \quad (3)$$

其中,C 为正确抽取的关系实例个数,M 为抽取出的关系实例总数,N 为标准结果集中的关系实例个数。

在开放式关系抽取兴起后,由于网络语料规模较大、噪声较多,难以简单地计算关系抽取结果的召回率,所以一般通过考察抽取准确率来评价系统性能。在考察抽取方法的实用性时,会加入运行时间和内存占用等指标。

3 传统的关系抽取方法

在 MUC 和 ACE 测评会议的推动下,关系抽取研究中出现了许多不同的方法。总体来看,这些方法可以分为两大类:基于知识工程的方法和基于机器学习的方法。基于知识工程的方法需要融合领域知识和语言学知识,通过人工编写规则集合,构造出特定模式,利用模式匹配的方式从文本中找到相应的关系实例。基于机器学习的方法一般将关系抽取问题转化为分类问题,通过特征工程选取有代表性的特征,利用不同的机器学习算法训练分类模型,最终通过训练出的分类模型判定实体对之间是否具有语义关系。也有的研究者通过聚类的方法解决关系抽取问题,取得了一定的效果。

3.1 基于知识工程的关系抽取

在 MUC-7 会议引入模板关系抽取任务后,研究者首先利用知识工程的方式解决关系抽取问题。Aone 等[9]通过人工编写抽取规则,从文本中抽取与规则匹配的关系实例。Fukumoto 等[10]利用实体之间的谓语信息判定两个实体之间的语义关系。Humphreys 等[11]将句法分析的结果作为输入,并人工标注实体之间的指代信息,利用复杂的句法规则识别实体之间的语义关系。

总体来说,基于知识工程的关系抽取方法能够在特定领域取得一定效果,但是存在三个缺陷:

(1)研究者需要在领域专家的指导下手工编写抽取规则集合,花费的时间较长。

(2)抽取系统的可移植性较差。当系统移植到其他领域时,需要重新编制抽取规则。

(3)当抽取规则集合较小时,规则的覆盖范围不够,抽取系统的召回率不高;当抽取规则集合比较复杂时,不同的规则之间容易产生冲突,导致抽取系统的准确率下降。

虽然基于知识工程的关系抽取方法存在缺陷,研究中也得出一些有益的启示:语义关系一般由连接实体的谓语表征;语义关系具有局部性,一般体现在包含两实体的窗口长度文本内;句法信息可以有效帮助识别出实体间的语义关系。这些启示促进了基于机器学习的关系抽取研究。

3.2 基于机器学习的关系抽取

2000 年,Miller 等[12]提出一个词汇化概率句法模型,利用宾州树库(Penn TreeBank)进行训练,在模板关系抽取任务中取得了较好的效果。Miller 等的研究表明利用机器学习方法解决关系抽取问题是可行的,利用词汇特征、句法特征训练模型,可以有效提升抽取系统的性能。基于机器学习的方法逐渐成为关系抽取研究的主流思路。根据人工参与和对标注语料的依赖程度不同,基于机器学习的关系抽取方法可以分为三大类:有监督的机器学习方法、半监督的机器学习方法、无监督的机器学习方法。

(1)有监督的机器学习方法

有监督的机器学习方法将二元关系抽取视为分类问题:

$$F(S) = \begin{cases} 1 & \text{如果 S 中的实体对具有某类语义关系} \\ -1 & \text{其余情况} \end{cases} \quad (4)$$

其中,S 为包含实体对的文本片段,F 为关系分类器,通常的模型有表决感知器(Voted Perception)和支持向量机(Support Vector Machines, SVMs)等。利用有监督机器学习方法解决关系抽取问题的一般流程是:人工标注训练样本得出正例和反例;通过选取特征集合,用已标注的正例和反例作为输入,训练出分类模型;用该分类模型对测试集合进行关系探测。根据关系实例的表示方式不同可以将有监督机器学习方法分为两类:基于特征向量的方法和基于核函数的方法。

Kambhatla[13]选取包含实体上下文、实体引用方式、依存关系、句法树在内的多种特征,实现了一个最大熵模型(Maximum Entropy Model)的关系分类器,在 ACE RDC2003 语料的 24 个子类关系抽取中获得 F 值

为 52.8% 的结果,表明多个层次的语言学特征能够提升关系抽取的效果。车万翔等[14]在 ACE RDC2004 语料上比较了 Winnow 和 SVM 分类器的效果,发现关系分类的性能相当。Zhou 等[15]在 Kambhatla 的实验基础上,新增了基本词组块特征,使用 WordNet 等语义资源,采用 SVM 作为分类器,获得了 F 值为 55.5% 的关系抽取结果,其实验结果表明实体类别特征对于关系抽取性能的提升最大。Jiang 等[16]将特征空间按照序列、句法和依存关系划分为不同的子空间,其实验表明特征子空间中的基本特征能有效提升关系抽取性能,而复杂特征带来的性能提升有限。董静等[17]将中文实体关系划分为包含与非包含关系,针对两类关系采用不用的句法特征,提高了汉语关系分类的性能。陈宇等[18]利用 DBN(Deep Belief Nets)模型进行关系抽取研究,证明字特征比词特征更适合中文关系抽取任务。

可以看出,特征工程(Feature Engineering)是基于特征向量的机器学习方法的核心。研究者通过启发式的方法选取特征集合,使用多层次的语言特征构造向量。目前很难找出适合关系抽取任务的新特征,因此一些研究者转向基于核函数的方法。

基于核函数的机器学习方法不需要人为构造显性的特征空间,而是直接以文本的字符串或者句法分析树结构作为输入,通过计算输入实例之间的相似度训练分类模型。在关系抽取任务中,树核函数的输入一般是句法分析结果及其各类变形。

Zelenko 等[19]利用浅层句法分析结果,用连接实体对的最小公共子树表征关系实例,通过计算两棵子树之间的核函数,训练 SVM 等分类器,在较小的新闻语料库中取得了较好的关系抽取效果。Culotta 等[20]改进 Zelenko 等的方法,利用依存关系句法树表示关系实例。通过添加词性、实体类型等特征,并在相似度计算时加入严格的匹配约束,从而使得关系抽取结果的准确率上升,但召回率下降。Bunescu 等[21]进一步改进,提出了实体对最短依存路径核函数,通过比较最短依存路径上相同节点的个数,计算核函数。其关系抽取结果同样遇到了召回率较低的问题。可以看出,基于句法树结构及其变形的树核函数方法,其最大的不足是相似度计算过程的匹配约束较为严格,导致最终抽取结果的召回率低。之后的研究一直围绕改善召回率来开展。为了解决这个问题,研究人员引入了卷积核函数。卷积核函数通过统计离散结构之间相同子结构的数目,计算两者的相似度。黄瑞红等[22]研究了卷积核方法对中文关系抽取的有效性,发现仅仅依靠最短依存路径核难以提高中文的实体关系抽取结果。Zhang 等[23]和 Zhou 等[24]利用实体对最短路径树,加入语义关系的不同层面特征,并综合考虑谓语上下文信息,利用卷积树核函数方法,有效地提升了关系抽取的性能。Qian 等[25]利用实体对的动态依存关系树,进一步改进了卷积树核函数方法的抽取性能。庄成龙等[26]在加入语义信息之外,对最短路径树进行裁剪,去掉修饰语冗余和并列冗余信息,较大地提升了关系分类效果。虞欢欢等[27]结合关系实例的结构化信息与实体语义信息,构造出合一句法和实体语义关系树,提高了中文语义关系抽取性能。刘克彬等[28]将改进后的语义序列核函数与 KNN 算法相结合构建关系分类器,在训练集较小时仍保持了较高的抽取精度。

基于核函数的方法可以利用文本的长距离特征,从而在理论上具有高维度的特征空间,实验结果也超过了基于特征向量的方法。但是由于核函数方法利用隐性方式表示特征,从而可能引入噪声信息,不利于判断特征有效性。同时,由于核函数的计算复杂度较高,分类器的训练和测试过程较慢,不适于处理大规模语料上的关系抽取任务。对基于特征向量和基于核函数的方法进行比较如表 1 所示:

表 1　基于特征向量和基于核函数方法的比较

方法	特征空间	特征表示	核心	处理速度
基于特征向量	词、词性序列、上下文、依存句法、句法树等	显式直观特征	特征工程	较快
基于核函数	树核、卷积树核等	隐式高维特征	核函数计算方式	较慢

综上,利用有监督的机器学习方法解决关系抽取问题时,需要启发式地结合各个层次的语言学特征,方能取得较好的实验结果。由于有监督的方法依赖标注语料库资源,难以适应自动处理的要求,因此研究在较少的人工参与和标注语料资源的情况下进行关系抽取,成为研究界的新热点。

(2)半监督的机器学习方法

在利用半监督的机器学习方法解决关系抽取问题时,主要采取基于自举(Bootstrapping)的思路:首先人工构造少量关系实例作为初始种子集合,然后利用模式学

习或者模型训练的方法,通过迭代过程,不断扩展该关系实例集合,最终获取足够规模的关系实例,完成关系抽取的任务。研究中一般由命名实体对作为关系种子。

Brin[29]基于自举方法,构建了 DIPRE(Dual Iterative Pattern Relation Expansion)系统。该系统利用少量实体关系对作为种子,通过不断迭代,自动从万维网页面中获取抽取模板和新的关系实例。Agichtein 等[30]设计实现了 Snowball 抽取系统,在 Brin 的工作基础上完善了关系的描述模式以及新抽取实例的可信度评价方式。Etzioni 等[31]构建了 KnowItAll 抽取系统,旨在从网络文本中抽取非特定领域的事实信息,通过多次改进,该系统获得了准确率为 90% 的抽取结果。Rosenfeld 等[32]和 Feldman 等[33]将迭代过程中获取的关系模式进行了泛化,弱化了匹配约束,同时加入命名实体识别信息,进一步提高了关系抽取效果。何婷婷等[34]在《人民日报》语料上测试了自举方法在中文关系抽取任务中的性能。李维刚等[35]利用网络挖掘方法获取了可信度较高的关系种子,增强了自举方法的性能。Xu 等[36-38]利用依存句法结构,实现了 DARE(Domain Adaptive Relation Extraction)关系抽取系统。该系统通过自底向上的方式构建大量规则模板,并利用背景知识[39]剔除错误和无效规则模板,提高了关系抽取的召回率。Zhu 等[40]利用马尔可夫逻辑网络(Markov Logic Networks, MLNs)改进了 Snowball 系统的模板评价方式,实现了 StatSnowball 抽取系统,在人际关系识别中取得不错的结果。Carlson 等[41]利用耦合半监督学习方法,在不同类别的抽取模板之间制定约束,有效减少了错误模板的产生,使得关系抽取准确率进一步提升。陈锦秀等[42]利用图模型表示关系实例集合,利用标注传递算法进行迭代,将标注实例的分类信息传播到临近的未标注样本上,在初始种子较少的情况下也能获取较好的抽取结果。

半监督机器学习方法可以有效地减少人工参与和对标注语料的依赖,并且能扩展到大规模文本的关系抽取任务,目前已被广泛采用。但是,自举方法在迭代过程中存在语义漂移(Semantic Drift)[43]的问题,影响抽取结果的准确率。因此,如何获取可信度较高的新关系实例和抽取模板,是目前研究的重点。

(3)无监督的机器学习方法

由于有监督和半监督机器学习方法需要事先确定关系类型,而实际上在大规模语料中,人们往往无法预知所有的实体关系类型。有些研究者基于聚类的思想,利用无监督机器学习的方法,尝试解决这个问题。

Hasegawa 等[44]通过将命名实体对之间的文本进行聚类,用聚类结果表示关系类别,使用聚类集合中词频最高的词作为关系描述词。在大规模新闻上的实验表明该方法是可行的。Stevenson[45]引入 WordNet 语义词典,改善了关系抽取模板聚类时的相似度计算过程。Zhang 等[46]利用浅层句法树表示关系实例,通过计算句法树之间的相似度,利用层次聚类算法进行聚类,该方法兼顾了低频实体对之间可能存在的语义关系。Rosenfeld 等[47]发现关系特征和实体特征的有效结合能大幅提高关系抽取的准确率,同时,在多种聚类算法中,单连通层次聚类算法是最优的。Davidov 等[48]利用 Google 搜索引擎,自动挖掘与特定概念词有关的实体和语义关系集合。Yan 等[49]将维基百科词条中的实体集合作为考虑对象,结合使用依存特征和浅层语法模板,通过模式聚类的方式,在大规模语料中抽取出对应实体所有的语义关系实例。Bollegala 等[50]从搜索引擎摘要中获取和聚合抽取模板,将模板聚类后发现由实体对代表的隐含语义关系。Bollegala 等[51]使用联合聚类(Co-clustering)算法,利用关系实例和关系模板的对偶性,提高了关系模板聚类效果,同时使用 L1 正则化 Logistics 回归模型,在关系模板聚类结果中筛选出代表性的抽取模板,使得关系抽取在准确率和召回率上都有所提高。

无监督的机器学习方法的应用场景一般是关系类型未事先确定,需要通过自动的方式将关系实例对应到正确的关系类型。无监督方法一般需要大规模语料作为支持,通过利用语料的冗余性,挖掘出可能的关系模式集合,然后确定关系名称。该方法的不足之处在于关系名称难以准确描述,同时,低频关系实例的召回率较低。

4 关系抽取技术发展趋势

传统的关系抽取研究一般面向特定领域,抽取特定实体之间有限的语义关系。近年来,关系抽取研究面向的领域逐渐转向海量网络文本,不少学者开始研究开放式关系抽取方法。TAC 会议将关系抽取划归为知识库构建的子领域,促进了基于 Distant Supervision

的关系抽取研究。同时,声明式关系抽取方法的发展,使得关系抽取技术进一步向实际应用转化。

4.1　开放式关系抽取

开放式关系抽取是信息抽取领域一种新研究范式,旨在从开放的网络文本中自动发现非限定类型的语义关系实例集合。大规模网络数据具有充分的冗余性,同时对处理过程的轻量化具有较高要求,因此开放式关系抽取具有与传统关系抽取不同的研究思路。

面向特定领域的关系抽取任务需要事先人为定义关系类型,同时系统移植性较差。为了解决这些问题,Sekine[52] 尝试了按需抽取(On - Demand Information Extraction)的思路,通过自动构造简单模板,完成非限定关系的关系抽取任务。Sekine 的工作表明了在开放文本环境下,利用浅层模式匹配的方法具有简便直观的优势。在 Shinyama 等[53] 研究的启发下,Banko 等正式提出了开放式信息抽取的研究思路,对于关系抽取的研究路线产生重要影响。Yates 等[54] 设计并实现了 TextRunner 系统,将动词作为关系名称,利用启发式规则和简单的句法特征训练分类器,在大规模网页上进行关系抽取,同时利用数据的冗余性判断抽取结果的可靠性。Wu 等[55] 实现了 WOE 系统,利用维基百科的 Infobox 中信息,提高了关系抽取效果。Etzioni 等[56]、Schmitz 等[57] 考虑到关系抽取在大规模文本上的可扩展性,在句法规则和词性信息的基础上实现 ReVerb 系统和 OLLIE 系统,抽取了以动词表示的实体关系,并减少了无信息量抽取和错误抽取的比例。

4.2　面向知识库构建的关系抽取

关系抽取是知识库自动构建[58] 的重要环节。研究者一般采取 Distant Supervision 的思路,即利用已有知识库蕴含的事实信息作为支撑,训练出抽取模型,在未标注的大规模语料上获取关系实例,从而补充已有知识库。Mintz 等[59] 利用 Freebase 丰富的实体关系信息训练分类器,从维基百科文章集合中抽取出新的关系实例。Krause 等[60] 则利用 Freebase 学习关系抽取规则。除了 Freebase 外,YAGO 也是 Distant Supervision 方法中常用的知识库[61,62]。

4.3　企业级应用中的关系抽取

在实际应用中,关系抽取是文本处理流程的重要环节,为保证结果的准确性,人工参与不可避免。以 IBM 公司为代表的企业一般采用声明式信息抽取(Declarative Information Extraction)的思路,通过机器辅助的方式帮助人们快速简便地完成抽取规则构建,从而提高处理效率。SystemT 系统[63] 利用声明语言 AQL 构建抽取规则,与基于 JAPE 规则[64] 进行抽取相比,取得了更好的效果。WizIE 系统[65] 包含了一个正则表达式生成模块,从而大大减少了人工编写抽取表达式的难度。由于 WizIE 的规则[66] 具有领域自适应性,所以可以处理大规模开放性文本,同时处理过程可以实现并行化。

5　结　语

关系抽取技术发展至今,在研究内容、方法路线和技术成熟度三个方面产生了较大的变化,呈现出清晰的发展脉络。

5.1　研究内容的变化

关系抽取研究内容和关系定义方式的转变如表 2 所示:

表 2　关系抽取研究内容和关系定义方式的转变

关键会议和研究趋势	研究内容	关系定义方式
MUC 会议	商业活动内容中的关系抽取	Location_of、Employee_of 和 Product_of 三种
ACE 会议	抽取机构关系、整体部分关系、人 - 社会关系等	事先定义的定义 7 大类
OpenIE 研究	开放领域的实体关系抽取	不事先指定,关系由抽取出的动词或者名词短语表示
TAC 会议	知识库构建中的槽填充任务	由背景知识库中已有的关系类型集合限定

从表 2 可知,在关系抽取的研究内容方面,发生了两个转变:

(1)关系抽取由限定领域转向开放领域。在 MUC - 7 会议上,关系抽取首先面向的是商业活动领域。ACE 会议的关系抽取语料范围有所扩展,而 OpenIE 的信息研究范式被提出后,关系抽取研究的领域扩展到了开放互联网领域。

(2)TAC KBP 会议将关系抽取研究定义为知识库领域的槽填充任务,使得关系抽取的研究范围兼顾了领域的开放性和体系性。

在关系定义方式上,由人工事先定义的有限类关系逐渐转变为未事先确定的开放类型关系。MUC - 7 会议上定义的关系类型是 Location_of、Employee_of 和 Product_of 三种,ACE 会议将关系类型扩展到 7 个大类 20 个左右的小类。OpenIE 的关系抽取研究中,用动词

或者名词短语表征关系,放宽了关系定义的约束,从而产生了更多精细的关系类型。TAC 会议将关系抽取视为知识库构建的子任务,从而使得待抽取的关系可以直接由知识库中已有的本体关系进行映射。关系抽取研究内容的深刻变化促进了关系抽取方法路线的转变。

5.2 方法路线的转变

在关系抽取的研究方法上,经历了"人工知识工程——机器学习——机器学习和模板匹配结合"的转变,如表 3 所示:

表 3 关系抽取方法路线的转变

关键会议和研究趋势	方法路线	代表性方法
MUC 会议	人工知识工程	基于人工规则
ACE 会议	有监督和半监督的机器学习	Bootstrapping
OpenIE 研究	模式匹配	基于词性标注的浅模式匹配
TAC 会议	机器学习 + 模式匹配	Distant Supervision

MUC - 7 会议上主要采取了人工编写规则的方式处理关系抽取问题,遇到了较大的困难。而 ACE 会议上的关系抽取研究则由机器学习方法主导,特别是 Bootstrapping 思路,对整个研究界产生了重大影响。随着 OpenIE 的兴起,研究者发现由于对标注语料的依赖,单纯的机器学习方法已经不能够很好地解决大规模海量数据集上的关系抽取任务。众多的研究表明,面向互联网开放领域的关系抽取中,机器学习和模板匹配具有各自的优点:基于机器学习的方法在标注语料资源充足的情况下能获取较好的效果,特别是将语义关系丰富的背景知识库作为训练样本后,能大大提升抽取模型的覆盖范围;基于模板规则的方法由于匹配精度高,处理速度快,能够适应大规模数据的实时处理要求。为了提升开放领域的关系抽取效果,一般将机器学习和模板匹配的方法结合,通过机器学习方法筛选出可信的模板集合,以实现对大规模文本的快速处理。面向知识库构建的关系抽取研究一般采取 Distant Supervision 思路,利用大规模知识库训练样本丰富的特性,从而获取更加完备和可信的抽取模板,同样也是"机器学习 + 模板匹配"的思路。

5.3 技术成熟度

MUC 和 ACE 测评会议提供了固定的测评语料,面向该测评语料的关系抽取系统能够获取 F_1 值为 75% 左右的结果。OpenIE 的研究面向的是广阔的互联网开放语料,对关系抽取结果的评价更依赖准确度或者可信度指标。卡内基梅隆大学研发的 NELL[67](Never - Ending Language Learning)系统已从互联网上抽取了 5 000 万的事实型信息,其中接近 200 万结果具有 95% 以上的可信度,占总体的 4% 左右。在实际的企业级应用中,关系抽取一般应用在具体的场景和领域中,仍需人工参与控制关系抽取精度,所以通过机器辅助的方式快速构建抽取规则,是性价比较高的方式。

综上所述,关系抽取是自然语言处理领域的重要研究方向,其研究内容已从限定领域、限定类型的关系分类转变为面向互联网开放领域的实体关系自动发现。随着关系抽取技术进一步实现自动化,将对海量信息处理、智能问答、知识库自动构建等领域产生积极推动,具有广阔的应用前景。

参考文献:

[1] Message Understanding Conference[EB/OL]. [2013 - 06 - 24]. http://en. wikipedia. org/wiki/Message _ Understanding _ Conference.

[2] MUC - 7 Information Extraction Task Definition[EB/OL]. [2013 - 06 - 24]. http://www. itl. nist. gov/iaui/894. 02/related_projects/muc/proceedings/ie_task. html.

[3] Automatic Content Extraction(ACE) Evaluation[EB/OL]. [2013 - 06 - 24]. http://www. itl. nist. gov/iad/mig//tests/ace/.

[4] The ACE 2007(ACE2007) Evaluation Plan[EB/OL]. [2013 - 06 - 24]. http://www. itl. nist. gov/iad/mig//tests/ace/ace07/doc/ace07 - evalplan. v1. 3a. pdf.

[5] Knowledge Base Population (KBP) 2013 [EB/OL]. [2013 - 06 - 24]. http://www. nist. gov/tac/2013/KBP/.

[6] Mintz M, Bills S, Snow R, et al. Distant Supervision for Relation Extraction Without Labeled Data[C]. In: *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*. Association for Computational Linguistics, 2009: 1003 - 1011.

[7] Banko M. Open Information Extraction for the Web[D]. University of Washington, 2009.

[8] Automatic Content Extraction 2008 Evaluation Plan (ACE08) [EB/OL]. [2013 - 08 - 24]. http://www. itl. nist. gov/iad/mig/tests/ace/2008/doc/ace08 - evalplan. v1. 2d. pdf.

[9] Aone C, Halverson L, Hampton T, et al. SRA: Description of the IE2 System Used for MUC - 7[C]. In: *Proceedings of the 7th Message Understanding Conference (MUC - 7)*. 1998.

[10] Fukumoto F, Shimohata M, Masui F, et al. Oki Electric Industry: Description of the Oki System as Used for MET－2[C]. In: *Proceedings of the 7th Message Understanding Conference*. 1998.

[11] Humphreys K, Gaizauskas R, Azzam S, et al. University of Sheffield: Description of the LaSIE－II System as Used for MUC－7 [C]. In: *Proceedings of the 7th Message Understanding Conference (MUC－7)*. 1998.

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

[13] Kambhatla N. Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations[C]. In: *Proceedings of the ACL* 2004 *on Interactive Poster and Demonstration Sessions*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004.

[14] 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. (Che Wanxiang, Liu Ting, Li Sheng. Automatic Entity Relation Extraction[J]. *Journal of Chinese Information Processing*, 2005,19(2):1－6.)

[15] Zhou G D, Su J, Zhang J, et al. Exploring Various Knowledge in Relation Extraction[C]. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics,2005:427－434.

[16] Jiang J, Zhai C X. A Systematic Exploration of the Feature Space for Relation Extraction[C]. In: *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL－HLT'07)*. 2007:113－120.

[17] 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. (Dong Jing, Sun Le, Feng Yuanyong, et al. Chinese Automatic Entity Relation Extraction[J]. *Journal of Chinese Information Processing*, 2007, 21(4): 80－85.)

[18] 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. (Chen Yu, Zheng Dequan, Zhao Tiejun. Chinese Relation Extraction Based on Deep Belief Nets[J]. *Journal of Software*, 2012,23(10): 2572－2585.)

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

[20] Culotta A, Sorensen J. Dependency Tree Kernels for Relation Extraction[C]. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004.

[21] Bunescu R C, Mooney R J. A Shortest Path Dependency Kernel for Relation Extraction[C]. In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics,2005: 724－731.

[22] 黄瑞红, 孙乐, 冯元勇, 等. 基于核方法的中文实体关系抽取研究[J]. *中文信息学报*, 2008, 22(5): 102－108. (Huang Ruihong, Sun Le, Feng Yuanyong, et al. A Study on Kernel－based Chinese Relation Extraction[J]. *Journal of Chinese Information Processing*, 2008, 22(5):102－108.)

[23] Zhang M, Zhang J, Su J, et al. A Composite Kernel to Extract Relations Between Entities with Both Flat and Structured Features [C]. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006: 825－832.

[24] Zhou G D, Zhang M, Ji D H, et al. Tree Kernel－based Relation Extraction with Context－Sensitive Structured Parse Tree Information[C]. In: *Proceedings of the* 2007 *Joint Conference on Empirical ethods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL－2007)*. 2007:728－736.

[25] Qian L H, Zhou G D, Kong F, et al. Exploiting Constituent Dependencies for Tree Kernel－based Semantic Relation Extraction [C]. In: *Proceedings of the 22nd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2008:697－704.

[26] 庄成龙, 钱龙华, 周国栋. 基于树核函数的实体语义关系抽取方法研究[J]. *中文信息学报*, 2009, 23(1): 3－9. (Zhuang Chenglong, Qian Longhua, Zhou Guodong. Research on Tree Kernel－based Entity Semantic Relation Extraction[J]. *Journal of Chinese Information Processing*, 2009, 23(1): 3－9.)

[27] 虞欢欢, 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. (Yu Huanhuan, Qian Longhua, Zhou Guodong, et al. Chinese Semantic Relation Extraction Based on Unified Syntactic and Entity Semantic Tree[J]. *Journal of Chinese Information Processing*, 2010,24(5):17－23.)

[28] 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. (Liu Kebin, Li Fang, Liu Lei, et al. Implementation of a Kernel－based Chinese Relation Extraction System[J]. *Journal of Computer Research and Development*, 2007, 44(8): 1406－1411.)

[29] Brin S. Extracting Patterns and Relations from the World Wide Web[C]. In: *Proceedings of International Workshop on the World Wide Web and Databases*. London, UK: Springer－Verlag, 1999:

172 – 183.

[30] Agichtein E, Gravano L. Snowball: Extracting Relations from Large Plain – text Collections[C]. In: *Proceedings of the* 5*th ACM Conference on Digital Libraries*. ACM, 2000:85 – 94.

[31] Etzioni O, Cafarella M, Downey D, et al. Unsupervised Named – entity Extraction from the Web: An Experimental Study[J]. *Artificial Intelligence*, 2005,165(1): 91 – 134.

[32] Rosenfeld B, Feldman R. URES: An Unsupervised Web Relation Extraction System[C]. In: *Proceedings of the COLING/ACL on Main Conference Poster Sessions*. Association for Computational Linguistics,2006:667 – 674.

[33] Feldman R, Rosenfeld B. Boosting Unsupervised Relation Extraction by Using NER[C]. In: *Proceedings of the* 2006 *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2006:473 – 481.

[34] 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 – 184,193. (He Tingting, Xu Chao, Li Jing, et al. Named Entity Relation Extraction Method Based on Seed Self – expansion[J]. *Computer Engineering*, 2006, 32(21):183 – 184,193.)

[35] 李维刚, 刘挺, 李生. 基于网络挖掘的实体关系元组自动获取[J]. *电子学报*, 2007,35(11): 2111 – 2116. (Li Weigang, Liu Ting, Li Sheng. Automated Entity Relation Tuple Extraction Using Web Mining[J]. *Acta Electronica Sinica*, 2007, 35(11): 2111 – 2116.)

[36] Xu F Y, Uszkoreit H, Li H. A Seed – driven Bottom – up Machine Learning Framework for Extracting Relations of Various Complexity [C]. In: *Proceedings of the* 45*th Annual Meeting of the Association of Computational Linguistics*. 2007: 584 – 591.

[37] Xu F Y. Bootstrapping Relation Extraction from Semantic Seeds [D]. Saarland University, 2008.

[38] Xu F Y, Uszkoreit H, Li H, et al. Adaptation of Relation Extraction Rules to New Domains[C]. In: *Proceedings of the Poster Session of the* 6*th International Conference on Language Resources and Evaluation*(*LREC*'08). 2008.

[39] Xu F Y, Uszkoreit H, Krause S, et al. Boosting Relation Extraction with Limited Closed – world Knowledge[C]. In: *Proceedings of the* 23*rd International Conference on Computational Linguistics*: *Posters*. Association for Computational Linguistics, 2010: 1354 – 1362.

[40] Zhu J, Nie Z, Liu X J, et al. StatSnowball: A Statistical Approach to Extracting Entity Relationships[C]. In: *Proceedings of the* 18*th International Conference on World Wide Web*. ACM, 2009:101 – 110.

[41] Carlson A, Betteridge J, Wang R C, et al. Coupled Semi – supervised Learning for Information Extraction[C]. In: *Proceedings of the* 3*rd ACM International Conference on Web Search and Data Mining*. ACM, 2010:101 – 110.

[42] 陈锦秀, 姬东鸿. 基于图的半监督关系抽取[J]. *软件学报*, 2008, 19(11): 2843 – 2852. (Chen Jinxiu, Ji Donghong. Graph – based Semi – Supervised Relation Extraction[J]. *Journal of Software*, 2008, 19(11):2843 – 2852.)

[43] Curran J R, Murphy T, Scholz B. Minimising Semantic Drift with Mutual Exclusion Bootstrapping[C]. In: *Proceedings of the* 10*th Conference of the Pacific Association for Computational Linguistics*. 2007:172 – 180.

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

[45] Stevenson M. An Unsupervised WordNet – based Algorithm for Relation Extraction[C]. In: *Proceedings of the* 4*th International Conference on Language Resources and Evaluation Workshop* "*Beyond Named Entity*: *Semantic Labelling for NLP Tasks*", Lisbon, Portugal. 2004.

[46] Zhang M, Su J, Wang D, et al. Discovering Relations Between Named Entities from a Large Raw Corpus Using Tree Similarity – based Clustering[C]. In: *Proceedings of the* 2*nd International Joint Conference on Natural Language Processing*(*IJCNLP*'05). Berlin, Heidelberg: Springer – Verlag, 2005: 378 – 389.

[47] Rosenfeld B, Feldman R. Clustering for Unsupervised Relation Identification[C]. In: *Proceedings of the* 16*th ACM Conference on Information and Knowledge Management*. ACM, 2007: 411 – 418.

[48] Davidov D, Rappoport A, Koppel M. Fully Unsupervised Discovery of Concept – specific Relationships by Web Mining[C]. In: *Proceedings of the* 45*th Annual Meeting of the Association of Computational Linguistics*. 2007:232 – 239.

[49] Yan Y, Okazaki N, Matsuo Y, et al. Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web [C]. In: *Proceedings of the Joint Conference of the* 47*th Annual Meeting of the ACL and the* 4*th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics,2009:1021 – 1029.

[50] Bollegala D T, Matsuo Y, Ishizuka M. Measuring the Similarity Between Implicit Semantic Relations from the Web[C]. In: *Proceedings of the* 18*th International Conference on World Wide Web*. ACM, 2009:651 – 660.

[51] Bollegala D T, Matsuo Y, Ishizuka M. Relational Duality: Unsupervised Extraction of Semantic Relations Between Entities on the Web[C]. In: *Proceedings of the* 19*th International Conference on World Wide Web*. ACM, 2010:151 – 160.

[52] Sekine S. On – Demand Information Extraction[C]. In: *Proceedings*

of the COLING/ACL on Main Conference Poster Sessions. Association for Computational Linguistics, 2006:731 – 738.

[53] Shinyama Y, Sekine S. Preemptive Information Extraction Using Unrestricted Relation Discovery[C]. In: *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 2006:304 – 311.

[54] Yates A, Cafarella M, Banko M, et al. TextRunner: Open Information Extraction on the Web[C]. In: *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics, 2007:25 – 26.

[55] Wu F, Weld D. Autonomously Semantifying Wikipedia[C]. In: *Proceedings of the* 16*th ACM Conference on Information and Knowledge Management* (*CIKM*'07), Lisbon, Portugal. 2007:41 – 50.

[56] Etzioni O, Fader A, Christensen J, et al. Open Information Extraction: The 2nd Generation[C]. In: *Proceedings of the* 22*nd International Joint Conference on Artificial Intelligence*. AAAI Press, 2011:3 – 10.

[57] Schmitz M, Bart R, Soderland S, et al. Open Language Learning for Information Extraction[C]. In: *Proceedings of the* 2012 *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012: 523 – 534.

[58] Ji H, Grishman R. Knowledge Base Population: Successful Approaches and Challenges[C]. In: *Proceedings of the* 49*th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 2011:1148 – 1158.

[59] Mintz M, Bills S, Snow R, et al. Distant Supervision for Relation Extraction Without Labeled Data[C]. In: *Proceedings of the Joint Conference of the* 47*th Annual Meeting of the ACL and the* 4*th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, 2009: 1003 – 1011.

[60] Krause S, Li H, Uszkoreit H, et al. Large – scale Learning of Relation – extraction Rules with Distant Supervision from the Web [C]. In: *Proceedings of the* 11*th International Semantic Web Conference*, Boston, MA, USA. Berlin, Heidelberg: Springer, 2012: 263 – 278.

[61] Akbik A, Visengeriyeva L, Herger P, et al. Unsupervised Discovery of Relations and Discriminative Extraction Patterns[C]. In: *Proceedings of COLING*. 2012:17 – 32.

[62] Nguyen T V T, Moschitti A. End – to – End Relation Extraction Using Distant Supervision from External Semantic Repositories [C]. In: *Proceedings of the* 49*th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 2011:277 – 282.

[63] Krishnamurthy R, Li Y Y, Raghavan S, et al. SystemT: A System for Declarative Information Extraction[J]. *ACM SIGMOD Record*, 2009, 37(4):7 – 13.

[64] Cunningham H, Maynard D, Tablan V, et al. JAPE: A Java Annotation Patterns Engine[OL]. [2013 – 06 – 24]. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid = B038120DE79C13635419187BFF58DFFF? doi = 10.1.1.32.3929&rep = rep1&type = pdf.

[65] Li Y Y, Chiticariu L, Yang H, et al. WizIE: A Best Practices Guided Development Environment for Information Extraction[C]. In: *Proceedings of the ACL* 2012 *System Demonstrations*. Association for Computational Linguistics, 2012:109 – 114.

[66] Chiticariu L, Krishnamurthy R, Li Y Y, et al. Domain Adaptation of Rule – based Annotators for Named – entity Recognition Tasks [C]. In: *Proceedings of the* 2010 *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010:1002 – 1012.

[67] Read the Web[EB/OL]. [2013 – 07 – 07]. http://rtw.ml.cmu.edu/rtw/.

（作者 E – mail:283306449@qq.com）