文章编号: 1006-2475(2018) 08-0021-07

中文实体关系抽取研究综述

武文雅,陈钰枫,徐金安,张玉洁

(北京交通大学计算机与信息技术学院,北京 100044)

摘要: 作为信息抽取任务中极为关键的一项子任务,实体关系抽取对于语义知识库的构建和知识图谱的发展都有着重要的意义。对于中文而言,语义关系更加复杂,实体关系抽取的作用也就愈加显著,因此,对中文实体关系抽取的研究方法进行详细考察极为必要。本文从实体关系抽取的产生和发展开始,对目前基于中文的实体关系抽取技术现状作了阐述;按照关系抽取方法对语料的依赖程度分为 4 类: 有监督的实体关系抽取、无监督的实体关系抽取、半监督的实体关系抽取和开放域的实体关系抽取,并对这 4 类抽取方法进行具体的分析和比较; 最后介绍深度学习在中文实体关系抽取上的应用成果和发展前景。

关键词: 中文实体关系抽取; 有监督方法; 无监督方法; 半监督方法; 开放域实体关系抽取方法; 深度学习

中图分类号: TP391 文献标识码: A doi: 10.3969/j.issn.1006-2475.2018.08.005

Review of Chinese Entity Relation Extraction

WU Wen-ya, CHEN Yu-feng, XU Jin-an, ZHANG Yu-jie

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Entity relation extraction is an important sub-task of information extraction. It is of great significance for the construction of semantic knowledge base and the development of knowledge graph. For Chinese, semantic relations are more complex, and the effect of entity relation extraction is more significant. So discussing the details of Chinese entity relation extraction methods is very necessary. From the beginning of the emergence and development of entity relation extraction, the current status of Chinese entity relation extraction technology is discussed. Relation extraction methods can be divided into four categories according to the degree of dependence on the corpus: entity supervised relation extraction, unsupervised relation extraction, semi-supervised relation extraction and open domain relation extraction. This paper analyzes and compares these four methods. Finally, the application results and development prospects of deep learning in Chinese entity relation extraction are introduced.

Key words: Chinese entity relation extraction; supervised method; unsupervised method; semi-supervised method; open domain entity relation extraction method; deep learning

0 引 言

20 世纪 90 年代中期以来,随着网络信息资源的日渐丰富、计算机速度的大幅度提高,主要以文字、图像等形式为依托的信息化时代强势到来。信息化时代的标志是信息爆发价值,如今信息化成为了时代发展的主流趋势,是前沿生产力的主要标志。随着信息时代的高速发展,信息数据呈现规模巨大、模态多样和高速增长等特征。在网络搜索过程中,当用户输入要查询的信息时,希望计算机能够从语义层面理解用户真实想要获取的信息,而不只是关键字的提取和搜索,这就迫切要求能快速、准确地获取用户真正所需信息的技术手段——信息抽取技术的进一步发展,以满足用户搜索的需求。比如说,当用户输入“英国伦敦”时,希望得到的是关于英国伦敦这座城市的多方面相关信息,如政治、经济、文化、地理位置、旅游景点、教育机制等,而不仅仅是简单的关于伦敦的关键字的句子提取。

收稿日期: 2017-10-31

基金项目: 国家自然科学基金资助项目(61473294, 61370130) ; 北京市自然科学基金资助项目(4172047) ; 中央高校基本科研业务费专项资金资助项目(2015JBM033)

作者简介: 武文雅(1995-) ,女,山西大同人,北京交通大学计算机与信息技术学院硕士研究生,研究方向: 实体关系抽取; 通信作者: 陈钰枫(1981-) ,女,福建南平人,副教授,博士,研究方向: 自然语言处理,人工智能; 徐金安(1970-) ,男,河南开封人,副教授,博士,研究方向: 自然语言处理,机器翻译; 张玉洁(1961-) ,女,河南安阳人,教授,博士,研究方向: 自然语言处理,机器翻译。

近年来,自然语言处理领域的研究者们开始致力于知识图谱构建的研究。知识图谱究其根本是一种语义网络图,通俗来讲,就是把多种信息按照一定的规则融合在一起而得到的关系网络。知识图谱是从“关系”的角度来分析问题的,为搜索提供了新思路:直接返回问题的答案,而不是返回包含关键词的文档序列。信息抽取则是知识图谱构建的关键一步。

信息抽取主要含有命名实体识别和实体关系抽取这 2 类子任务。命名实体识别指的是从自然语言文本中识别出实体类、时间类和数字类 3 大类,以及人名、机构名、地名、时间等 7 小类命名实体[1]。命名实体识别准确率的提高可以促进信息提取、语篇理解、句法分析以及机器翻译等任务的发展,对自然语言处理技术产业化发挥着奠基性的作用,但是实体识别得到的只是以离散形式存在的实体,并不能有效地反映命名实体之间的关系,而实体关系抽取就是用来处理这个问题的方法。实体关系抽取是从自然语言文本中辨别出 2 个实体间所存在的语义关系,例如,对于句子“李克强在阿斯塔纳会见阿富汗首席执行官阿卜杜拉。”中的实体“李克强”和“阿卜杜拉”之间存在着“会见”关系,同时,实体“阿富汗”和“阿卜杜拉”间拥有“首席执行官”的关系。作为自然语言处理的重要任务之一,实体关系抽取为海量信息处理、中文信息检索、知识库自动构建、机器翻译和自动文摘等众多自然语言处理任务提供了重要的技术支持。

1 实体关系抽取的产生与发展

1.1 实体关系抽取的产生

美国国防高级研究计划委员会(DARPA)资助的 MUC 会议鼓励关于信息抽取新方法的提出[2]。1995 年举办了 MUC-6 会议,前面几届会议都聚焦在“信息提取”任务上:分析自由文本,识别某种特定类型的事件,并使用每个事件的信息去填充数据库模板。随着前 5 次 MUCs 的开展,任务和模板变得越来越复杂。NYU 和 NRaD 合作提出了命名实体识别(NER)、指代(Coreference)、模板元素(Template Elements)和场景模板(Scenario Templates)[2] 这 4 项任务。

1998 年最后一次 MUC-7 会议在 MUC-6 会议任务的基础上初次提出了关系抽取(模板关系,Template Relation)任务,是用模板关系来进行描述的[3]。MUC-7 会议的语料是与飞机失事事件(airplane crashes)和航天器发射事件(rocket missile launches)相关的新闻报道,主要包含 LOCATION_OF、EMPLOYEE_OF 和 PRODUCT_OF[4] 这 3 种实体关系类别。

1.2 实体关系抽取的发展

MUC 会议一共举办了 7 届,进入 21 世纪后,美国国家标准技术研究所(NIST)组织的自动内容抽取(ACE)评测会议成为信息抽取研究进一步发展的主要动力[5]。自动内容抽取(ACE)评测会议是 21 世纪初期继 MUC 会议之后,文本分析会议(TAC)之前的研究先进信息抽取技术的会议。该评测会议将实体关系识别作为一项重要的评测任务进行发布[5]。

ACE 的实体关系语料是语言资源联盟(LDC)供给的,语种已由单一的英文扩展到了阿拉伯语、西班牙语和中文。中文的数据是由哈工大自然语言处理实验室标注的,数据内容涉及广播新闻、新闻专线和网络会话。ACE 的实体关系语料的语种数量和数据规模在 MUC 的基础上都有了大幅度的扩展。ACE 2008 的关系抽取任务共定义了组织机构—附属、部分—整体、人—社会等 7 个大类的实体关系,细分为使用者—拥有者—发明人—制造人、公民—居民—宗教人士—种族人士、组织—位置等 18 个子类的实体关系。ACE 评测会议给实体关系抽取研究提供了新的发展平台[5]。从 2009 年开始,ACE 被归入文本分析会议(TAC),成为了 Knowledge Base Population 工程中不可缺少的一部分[6]。

除了 MUC 和 ACE 会议之外,语义评估(Semantic Evaluation, SemEval)会议也是自然语言处理领域中一个极具影响力的评测会议。该会议聚焦于句子级单元间的彼此联系(例如语义角色标注)、语句间的联系(例如指代)和人们所说的自然语言(语义关系和情感分析)。SemEval-2007 的评测任务 4 中设置了 7 种常用名词和名词短语间的实体关系,在 SemEval-2010 第 8 项任务中将实体关系类型扩充到了 9 种:Component-Whole、Instrument-Agency、Member-Collection、Cause-Effect、Entity-Destination、Content-Container、Message-Topic、Product-Producer 和 Entity-Origin。在 2010 年的评测中掀起了普通名词和名词短语间实体关系抽取研究的新高潮[7]。

MUC、ACE、SemEval 评测会议所用的实体关系语料都是事先标注好的,即由领域专家制定好关系类型体系,然后对大规模文本进行人工逐个判断。这样的方法耗时耗力,成本极高,同时不利于扩展语料类型。近年来,开放域实体关系抽取方法逐渐受到关注,相比传统实体关系抽取来说,在语料方面它解决了语料获取困难的问题。Wikipedia、HowNet、WordNet 和 Freebase 等涵盖大规模事实性信息的知识库为标注语料的获取提供了有效的数据支持。与传统

的人工标注语料方法相比较,基于 Web 开放语料的规模更宏大,涉及的领域更广阔,囊括的关系类型也更丰富[8]。

2　中文实体关系抽取的研究现状

在当今时代中,中文在全球的使用越来越广泛,因此对中文实体关系抽取的研究也日趋紧迫。根据输入数据是否有标签,即语料中的实体关系是否被标注出来,本文把中文实体关系抽取方法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域抽取方法[9],下面对这几种方法分别进行相关介绍。

2.1　有监督的中文实体关系抽取

有监督的实体关系抽取方法是最早开始使用的,也是发展最快的方法。在这种学习方法中,关系抽取常被当作分类问题来解决。关系抽取所依赖的方法基本可以归纳为: 基于模式匹配的方法和基于机器学习的方法[11]。其根据关系实例的表示方式不同分为基于特征的方法和基于核函数的方法[10]。

基于模式匹配的关系抽取方法需要领域专家和语言学家互相合作,运用语言学知识和专业领域知识构造出基于词语、词性或语义的模式集合。通过将预处理后的语言片段和模式进行匹配来实现关系抽取,如果两者相匹配,则可以说该语句拥有相应模式的关系属性。这种方案的关键之处在于关系模式的确立,关系模式的建立需要语言学家对领域专业知识通达,穷举所有可能的关系表达,人工罗列关系模式。限于语言学家对专业知识的了解,该方法既费时费力,又不可避免地出现错误; 同时领域自适应性能极差,当出现新领域语料时,需要语言学家重新列举关系模式,研究者针对此问题提出了一些经过实验证明可行的解决方法[11]。

Appelt 等[12]在 MUC-6 会议上阐述了 FASTUS 抽取系统,并提出了“宏”这一新概念,用一般通用形式来构建领域规则。若想迅速构建不同专业领域的关系模式体系只需要重置相应“宏”中的参数。Yangarber 等[13]在 MUC-7 会议上展示的 Proteus 抽取系统中融合了样本泛化的关系抽取模式方法,一定程度上提高了模式构建的领域自适应能力。周诗咏[14]提出了一种融合语义模式匹配的实体关系抽取模型 SPMREM,在分析文本语义结构的基础上提出一种字符串匹配方法,并结合相似密度方法对关系模式进行聚类,提取关系模式集,实验表明该方法能高效地从语料中抽取相关的关系模式。

基于模式匹配的方法已有了一定的成效,同时基于特征向量的关系抽取目前也已经取得了不可忽视的成果。若想改进基于特征向量的方法则需要在有效特征的抽取与集成上下功夫。准确地获取词法、句法、语义等特征,并把它们融合在一起,这才是特征向量方法进步的根源。中文实体关系抽取的结果通常采用 F 值来进行评价,计算方法如下[15]:

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

与 F 值计算相关的准确率(Precision) 和召回率(Recall) 的计算公式如下所示[15]:

$$\text{Precision} = \frac{\text{某类被正确分类的关系实例个数}}{\text{被判定为某类的关系实例个数}}$$

$$\text{Recall} = \frac{\text{某类被正确分类的关系实例个数}}{\text{测试集中某类关系实例个数}}$$

在基于特征向量的中文实体关系抽取研究方面,车万翔等[15]运用 Winnow 和 SVM 算法,谨慎研究比较发现,当选取每个实体的周围 2 个词为特征时,抽取效果达到最佳。在 ACE RDC 2004 语料上实验,加权平均 F 值分别达到 73.08% 和 73.27%。在不同的语料上,通常会存在不同的最优特征向量,例如,在 ACE2004 的语料上,车万翔等[15]实验得出实体周围 2 个词为最优特征,但是在微博新闻语料上,当选取实体周围 3 个词为特征时,效果要优于 2 个词的情况。黄鑫等[16]提取了语料词语、实体和语法的基本特征后,将其进行特征组合,在 ACE RDC 2005 中文语料上进行实验,结果表明组合特征的性能比单独特征的性能更好,F 值平均提高了 2.0%。一般而言,组合特征的效果由于融合了多种特征,在效果上有一定的提升,除非组合特征过于贴合训练集,过拟合现象严重,会导致 F 值不升反降。郭喜跃等[17]在词法特征和实体原始特征的基础上加入了依存句法关系、核心谓词和语义角色标柱等特征,实验结果表明加入的多种特征对关系抽取准确率的提高颇有帮助。依存句法信息和语义角色标注等信息的加入使得特征向量更加饱满,特征更具代表性,极大程度上提高了关系抽取的性能。

有监督关系抽取方法的另一个主流方法是基于核函数的方法。早期使用较多的是序列核函数,它一方面拥有较好的复合性能,另一方面考虑了特征间的顺序和结构信息。在序列核函数的基础上加入语义信息可以拓展算法应用范围。将多种核函数融合在一起联合抽取实体关系会充分发挥各种核函数的优势,实验证明其结果提升明显。虞欢欢等[18]构造了句法和语义关系树,方法是将实体语义信息加入关系实例的结构化信息中,不仅可以获得结构化信息,还可以获取实体语义信息。在 ACE RDC 2005 中文语

料上进行的实验表明,构造实体语义结构树方法能提高 F 值,同时也说明规则化的结构句法信息和语义信息相结合可以增强有效特征。Zhou 等[19] 提出了一种基于树核的语义信息抽取方法,利用解析树和实体对构造丰富的语义关系结构,来综合句法和语义信息,在 ACE 语料上的实验结果表明这种树核方法在当时处于世界领先水平。王敏[20] 把基于特征向量的平面核融合到了基于句法分析树的结构核中,这种多核融合的方法使得关系抽取性能得到了提升。陈鹏等[21-22] 详细考究了特定领域信息的特征,构造出含有语义关系的领域知识树,并将其应用到领域信息的句法树中。在旅游相关领域的语料上进行关系抽取实验,结果表明由于引入了语义知识该方法优化了关系抽取性能。郭剑毅等[23] 针对传统径向基核函数的训练矩阵中所有元素接近于零不利于分类的问题,提出了一种向量离散化的训练矩阵,将改进的径向基核函数融合多项式函数及卷积树核函数进行实验,实验结果证实了相对单一核函数,改良的多核融合方法性能更优。

基于核函数方法在运算速度上有一定的弊端,尽管如此,研究者还是希望通过对核函数的进一步研究来获得关系抽取的进步。

2.2 半监督的中文实体关系抽取

运用半监督的方法进行中文实体关系抽取,只需要少量的标注数据,因此,当需要处理标注语料较少的实体关系抽取任务时,可以选用半监督的方法。

自举方法(Bootstrapping)、协同训练(Co-training)和标注传播(Label propagation)方法是目前在关系抽取任务中经常使用的半监督方法,以下分别进行介绍。

Brin[24] 首先使用了基于 Bootstrapping 的半监督方法进行实体关系抽取。该方法首先需要确立关系种子类型,接着从包含种子的上下文中总结关系模式从而寻找更多的关系种子实例以便于扩充种子集合,最后迭代得到领域关系实例和序列模式。余丽等[25] 运用 Bootstrapping 方法在地理领域的语料上分析词语的特征,比如词性、位置、距离,根据这些特征来提取表示实体关系的关系指示词。该方法能自动挖掘自然语言的部分词法特征,避免对大规模标注语料的依赖,适合用于缺乏大量标注语料的关系抽取任务。

基于 Bootstrapping 的方法对初始关系种子的质量要求较高,如果初始关系种子选择不恰当,会对种子集合的扩展有影响。当领域发生迁移时需要重新确立序列模式并且重新构建高质量的关系种子。

基于协同训练思想的 BootProject 方法被 Zhang 提出用来进行半监督语义关系分类[26]。BootProject 方法是从一个大的特征集合中任意抽取出含有合适数目的特征子集当作一个窗口,反复此进程获得多个窗口,运用开始少量的种子集合语料训练分类器,对实例进行分类,以此找出有代表性的关系实例,投入种子集合中以便下一次的种子集拓展。初始种子集大小与结果准确率有关,在一定范围内,它们成正相关。张一昌[27] 将协同训练关系抽取方法和核函数融合在一起,F 值提高了 0.05%,同时,他还把 Word-embedding 应用于协同训练关系抽取中,使 F 值有了 0.1% 的提升。这个方法涉及的 2 个问题是,怎样抽取理想的初始种子集以及怎样减少迭代过程中的错误实例数量。

标注传播算法是一种基于图的半监督学习方法,它的目的是训练计算机从半结构化或者非结构化的文本中自动识别出实体对之间存在的关系。该方法的特别之处在于利用图策略建立关系抽取模型,图上的节点表示样本实例,图上边的权重表示样本实例之间的距离,关系抽取任务就此转化成为根据该图估计一个满足全局一致性假设的标注函数,这种任务转化的思想为中文实体关系抽取任务提供了新的解决思路。当标注数据较为缺乏时,标注传播算法在中文实体关系抽取任务中往往可以取得远高于有监督方法(SVM,NB,RNN)的抽取准确率[28]。这是因为标注传播算法可以借助图模型来平滑无标签样本的标签信息。也就是说,在半监督学习方法中,无标签样本的标签信息同时由与其相邻的有标签样本和无标签样本来决定。然而在有监督方法中,无标签样本的标签信息仅仅取决于与之相邻的有标签样本。郝建柏[29] 提出基于局部线性嵌入算法构建图的标签传递算法,该算法中的图比传统图更容易使用,分类精度更高,在实验中的结果也证实了这一点。该方法的缺点是占用更多的存储空间,运行时间也比较长。

对于半监督学习,共同存在的问题是初始种子集的选取,以及如何缓解迭代过程中的噪音干扰等语义漂移问题。当然,进一步探索新的半监督学习方法是提高半监督学习抽取性能的有效手段。

2.3 无监督的中文实体关系抽取

在没有标注数据的情况下,研究者们使用无监督方法进行中文实体关系抽取,主要包括实体对聚类和关系指示词选择 2 部分。具体做法是首先将上下文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一类关系。

Hasegawa 等[30] 在 ACL 会议上首次使用了无监

督的关系抽取方法,该方法识别出实体对的类型,把共同出现次数多于一定阈值的实体对作为潜在的语义关系,并且计算实体对间的词汇相似度对其进行聚类,最后根据经验给聚类的实体对冠上合适的关系名称。实验中使用这个方法发现的公司实体对之间的语义关系 F 值高达 0.75,实验证明这种无指导的方法效果较为明显。Rink 等[31] 基于产生式模型构建了无监督实体关系抽取框架,实现了医学专业领域中实体关系的有效抽取,这一在特定领域中进行无监督的关系抽取方法的应用,在一定程度上促进了关系抽取产业化的进一步发展。孙勇亮[32] 采用密度聚类算法,在无监督实体关系抽取任务中获得了不错的结果,实验表明优化聚类算法对无监督关系抽取性能的提升有着重要的作用。王晶[33] 提出了一种语料相关的提取特征算法,其中考虑到了启发式规则,并且根据数据集特征孕育出一种新的聚类算法,在大规模网络文本中进行实验,表明该方法在关系抽取任务上有效果。施琦[34] 使用了一种弹性上下文窗口代替传统固定窗口大小的模式来进行特征词的选取,并且充分利用互信息计算特征词权值同时融入了改进的 k-means 算法,在网络文本上的实验表明,这些改进都可以使关系抽取的精度提高。

使用无监督的方法进行实体关系抽取不需要预先定义实体关系类型体系,领域适应性强,在处理大规模网络文本数据时极具优势,改进方法主要在于选择合适的特征和优化聚类算法。无监督的实体关系抽取需要预先确定聚类阈值,这是该方法的难点,同时,无监督的实体关系抽取尚缺乏客观的评价标准。

2.4　开放域中文实体关系抽取

近年来,专家学者们提出了一种针对开放领域实体关系抽取的 Open Information Extraction(Open IE) 方法[35-37],不需要人工标注语料,也不需要事先知道抽取哪些实体关系。它的目标是自动将自然语言句子转换为有意义的事实性命题。例如,输入句子“莫言,山东高密人,是中国历史上第一位获得诺贝尔文学奖的作家。”输出命题: 莫言,是,山东高密人; 莫言,是,作家; 莫言,是,中国人; 莫言,第一位获得,诺贝尔文学奖。通过对输出命题的分析,可以得到很多有效的信息。在海量网络文本数据中,可以通过开放式关系抽取快速地从中提取大量的实体关系三元组。例如,从“北宋有名的诗人范仲淹政绩突出,文学成就卓越。”中抽取出(北宋,诗人,范仲淹) 这个关系三元组。“北宋”和“范仲淹”这 2 个实体的关系用句子当中的名词“诗人”来描述。当然,通常可以用句子当中的名词、动词或者名词性短语来描述实体关系。

Open IE 方法的难点主要在于复杂句子的处理和关系短语语义的归一化,不是所有的句子都能很容易地找到正确命题。对于开放域实体关系抽取可用的技术有句法模式学习、自学习技术、句子分解技术、Clustering 和 Inference Rule Discovery 等。针对复杂句子的处理,Corro 和 Gemulla 等[38] 提出了 Clause IE,它根据语言语法规律定义了 7 种简单句子模式和一系列句子分解规则,将复杂句子分解为简单句,借助化繁为简的方法,使复杂句子转化为简单句来进行处理。对于语义的归一化,目前的解决方案是计算不同关系短语之间的相似度来识别表达相同语义的关系短语,代表性模型有 Topic Model[39]、Random Walk[40]。

Washington 大学的人工智能研究组在开放式实体关系抽取领域作出了很大贡献。TextRunner[41]、WOE[42] 等系统都是其开发用于开放域关系抽取研究的。

目前,对于中文开放域实体关系抽取也有了一定的进展。秦兵等[43] 在大规模的网络文本上进行了无监督的实体关系抽取。通过观察,其首先利用实体间的距离和关系指示词的位置限制得到大量的候选关系三元组,接着使用基于规则的方法提取能正确表示实体间关系的关系指示词,最后通过对错误三元组进行分析,构建合适的句式规则,对其过滤得到精确度较高的实体关系三元组,可用于充实文本知识库。郭喜跃[44] 在百科类开放领域文本上使用弱监督方法获取了高质量的关系三元组,其在借助于百度百科信息框得到标注语料的同时对其进行筛选和合并,这种做法使初始语料的质量有了进一步的提高。通过对初始语料的加工,其整体 F 值达到了 79.27%。针对存在多元实体关系的抽取问题,李颖等[45] 运用依存关系分析来抽取多元实体关系,在百度百科数据集上的抽取准确率可达 81%。

开放式实体关系抽取还存在很大的进步空间。其一,由于数据来源的不统一,实体关系抽取结果评价体系还没有达成一致标准; 其二,当前大部分实验都是在数据进行大量清洗之后的干净数据上进行的,数据真实性难免会有所下降。如何在真实网络数据上进行关系抽取是要继续研究的重点问题之一。

2.5　4 种实体关系抽取方法的总结

针对中文的实体关系抽取任务,上面所介绍的 4 种方法各有优劣,表 1 对上述的 4 种方法作了总结。从实现方法、泛化能力、对语料标注的依赖程度和性能提升方法等方面对这 4 类实体关系抽取方法进行了详细的比较。

表 1 实体关系抽取方法总结

实体关系抽取方法	实现方法	领域泛化能力	对标注语料依赖程度	优点	缺点	性能提升方法
有监督	分类	弱	强	能充分利用先验知识; 通过有效特征的选择,往往能取得更好的性能	需要大量标注数据; 需要预先确定抽取的关系; 领域自适应性不强	改进规则,构造代表性特征,优化核函数
半监督	分类	中	中	需要较少的标注语料; 适合开放/Web 环境下的关系抽取	对结果要进行大量分析和后处理; 性能不及有监督方法	优化模式扩展和噪声过滤方法
无监督	聚类	强	弱	不需要标注语料; 适合处理大规模无标注语料	需要事先确定聚类阈值	增加特征,优化聚类算法
开放域	分类	强	弱	不需要标注语料; 不需要事先知道抽取的关系; 适合处理大规模网络数据	对抽取的结果要进行大量后处理; 缺乏客观的评价标准	优化文本的噪声过滤算法

3 实体关系抽取中深度学习的应用

21 世纪以来,深度学习被广泛运用于自然语言处理任务。近十多年来,深度学习也开始在实体关系抽取领域取得成果。

Socher 等[47] 提出运用递归神经网络来应对关系抽取任务,该方法考虑了句子的句法结构信息,但是无法考虑到实体对的位置和语义信息。Zeng 等[48] 应用卷积神经网络解决关系抽取任务,向卷积神经网络输入词向量和词位置向量,之后通过卷积、池化和非线性变换得到句子表示。由于考虑到了实体的位置向量和其他相关词汇特征,句子的实体信息也同步被关注到了。Miwa 等[49] 提出了一种双向 LSTM 和树形 LSTM 模型相结合的方法。该方法运用这 2 种网络的同时对实体和句子进行建模,取得了较好的效果。

除了在有标注语料上的研究,关系抽取在纯文本上的研究也有了一定的进展。Lin 等[50] 提出了一种在纯文本中进行关系抽取的方法。他们引入了一种多语言的神经关系抽取框架,在单语文本中采用单语注意机制,并且提出跨语言注意机制来考虑跨语言文本信息的一致性和互补性。Lin 等[51] 运用基于句子级别注意力机制的神经网络模型解决了实体对对应的噪音句子问题,使得模型能利用所有有效句子进行学习,通过实验发现此方法有效地控制了噪音句子的影响,使得关系抽取效果得到了提升。

以上都是深度学习在英文语料上的应用,在中文研究方面,由于标注语料的短缺,深度学习在实体关系抽取领域上的应用相对于在其它自然语言处理任务上的应用来说较少。孙建东等[52] 在 COAE2016 的数据集上使用了卷积神经网络模型,但是由于数据集较少的原因,效果比 SVM 方法的 F1 值低近 10%。

深度学习中的卷积神经网络、循环神经网络、LSTM 网络等架构在自然语言处理领域的应用极其广泛,并且取得了良好的效果。如果能将深度学习方法运用于缺乏标注的数据集上,那么中文实体关系抽取研究将取得进一步的成果。

4 结束语

尽管实体关系抽取在一定程度上已经取得了不菲的成绩,但是在中文语料上的研究成果还有待提高。有监督的实体关系抽取方法将关系抽取任务当作分类任务,在标注语料上提取有效的特征,训练分类器来预测实体关系,特征的选取对于实验结果往往起着至关重要的作用。无监督实体关系抽取领域移植性强,适合处理大规模无结构的网络文本数据。半监督实体关系抽取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。开放式实体关系抽取不需要事先定义好关系类型,直接用句子中的关系指示词来表示实体关系,具有广阔的发展前景。近年来快速发展起来的 Deep Learning 方法在中文实体关系抽取任务上还没有大量的应用,主要是因为标注数据集匮乏,如果无监督的中文实体关系抽取在大规模网络文本上有了飞跃性的发展,那么 Deep Learning 也将在此领域大放异彩。无论运用哪一种方法,不断的技术更新是不可缺少的,只有克服相应的问题,找出解决办法或者替代方法,基于中文的实体关系抽取研究才能取得长足的进展。

参考文献:

[1] 百度百科. 命名实体识别[EB/OL]. https://baike.baidu.com/item/%E5%91%BD%E5%90%8D%E5%AE%9E%E4%BD%93%E8%AF%86%E5%88%AB/6968430, 2017-10-28.

[2] NYU. MUC-6 [EB/OL]. http://cs.nyu.edu/cs/faculty/grishman/muc6.html, 1996-04-25.

[3] Chinchor N, Marsh E. Muc-7 information extraction task definition [C] // Proceedings of the 7th Message Understanding Conference(MUC-7). 1998:359-367.

[4] Chinchor N A. Overview of MUC-7/MET-2 [EB/OL]. https://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html, 2005-05-08.

[5] Linguistic Data Consortium. The Automatic Content Extraction(ACE) Projects [EB/OL]. http://www.ldc.upenn.edu/Projects/ACE/, 2007-01-11

[6] McNamee P, Dang H T, Simgpson H, et al. An evaluation of technologies for knowledge base population [C] // Proceedings of the 7th International Language Resources and Evaluation Conference. 2010:369-372.

[7] Hendrickx I, Kim S N, Kozareva Z, et al. Semeval-2010 task 8: Multi-way classification of semantic relation between pairs of nominals [C] // Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. 2009:94-99.

[8] 刘绍毓,李弼程,郭志刚,等. 实体关系抽取研究综述[J]. 信息工程大学学报, 2016,17(5):541-547.

[9] 张传岩. Web 实体活动与实体关系抽取研究[D]. 济南: 山东大学, 2012.

[10] 李天颍,刘磷,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.

[11] 徐健,张智雄,吴振新. 实体关系抽取的技术方法综述[J]. 现代图书情报技术, 2008,24(8):18-23.

[12] Appelt D E, Hobbs J R, Bear J, et al. SRI international FASTUS system: MUC-6 test results and analysis [C] // Proceedings of the 6th Message Understanding Conference (MUC-6). 1995:237-248.

[13] Yangarber R, Grishman R. NYU: Description of the Proteus/PET system as used for MUC-7 ST [C] // Proceedings of the 7th Message Understanding Conference. 1998:1-7.

[14] 周诗咏. Web 环境下基于语义模式匹配的实体关系提取方法的研究[D]. 沈阳: 东北大学, 2009.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Zhou Guodong, Qian Longhua, Fan Jianxi. Tree kernel-based semantic relation extraction with rich syntactic and semantic information [J]. Information Sciences, 2010,180(8):313-1325.

[20] 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, 2011.

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

[22] 陈鹏. 基于多核融合的中文领域实体关系抽取研究[D]. 昆明: 昆明理工大学, 2014.

[23] 郭剑毅,陈鹏,余正涛,等. 基于多核融合的中文领域实体关系抽取[J]. 中文信息学报, 2016,30(1):24-29.

[24] Brin S. Extracting patterns and relations from the World Wide Web [C] // WebDB Workshop at the 6th International Conference on Extended Database Technology. 1999:172-183.

[25] Yu Li, Lu Feng, Liu Xiliang. A bootstrapping based approach for open geo-entity relation extraction [J]. Acta Geodaeticaet Cartographica Sinica, 2016,45(5):616-622.

[26] Zhang Zhu. Weakly-supervised relation classification for information extraction [C] // Proceedings of the 13th ACM International Conference on Information and Knowledge Management. 2004:581-588.

[27] 张一昌. 基于 co-training 与核函数的关系抽取技术研究[D]. 北京: 北京邮电大学, 2015.

[28] 罗斌,唐红艳,王志豪,等. 基于图的微博广告文本识别[J]. 厦门大学学报(自然科学版), 2017,56(5):724-728.

[29] 郝建柏. 基于图的半监督学习模型研究与分类器设计[D]. 合肥: 中国科学技术大学, 2009.

[30] 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: Article No. 415.

[31] 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.

[32] 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.

[33] 王晶. 无监督的中文实体关系抽取研究[D]. 上海: 华东师范大学, 2012.

[34] 施琦. 无监督中文实体关系抽取研究[D]. 北京: 中国地质大学(北京), 2015.

[35] Kang Tian, Zhang Shaodian, Tang Youlan, et al. EliIE: An open-source information extraction system for clinical trial eligibility criteria [J]. Journal of the American Medical Informatics Association, 2017:24(6):1062-1071..

[36] Imani M. Evaluating Open Relation Extraction over Conversational Texts [D]. University of British Columbia, 2014.

[37] Wang M, Li L, Huang F. Semi-supervised Chinese open entity relation extraction [C] // IEEE International Conference on Cloud Computing and Intelligence Systems. 2015:415-420.

(下转第 34 页)

gence, 2016,48(C):59-71.
[8] 黄光球,赵魏娟,陆秋琴. 求解大规模优化问题的可全局收敛蝙蝠算法[J]. 计算机应用研究, 2013,30(5):1323-1328.
[9] Khan K, Nikov A, Sahai A. A fuzzy bat clustering method for ergonomic screening of office workplaces[C]// The 3rd International Conference on Software, Services and Semantic Technologies S3T. 2011:59-66.
[10] Yang Xinshe. Bat algorithm for multi-objective optimisation[J]. International Journal of Bio-inspired Computation, 2012,3(5):267-274.
[11] Komarasamy G, Wahi A. An optimized K-means clustering technique using bat algorithm[J]. European Journal of Scientific Research, 2012,84(2):263-273.
[12] Lin J H, Chou Chaowei, Yang C H. A chaotic Levy flight bat algorithm for parameter estimation in nonlinear dynamic biological systems[J]. Journal of Computer and Information Technology, 2012,2(2):56-63.
[13] Nakamura R Y M, Pereira L A M, Costa K A. BBA: A binary bat algorithm for feature selection[C]// Proceedings of 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images. 2012:291-297.
[14] Xie Jian, Zhou Yongquan, Chen Huan. A novel bat algorithm based on differential operator and Sévy flights trajectory[J]. Computational Intelligence & Neuroscience, 2013.
[15] Zhang Jiawei, Wang Gaige. Image matching using a bat algorithm with mutation[J]. Applied Mechanics and Materials, 2012,203(1):88-93.
[16] Kacem A, Lallemand C, Giraud N. A small-world network model for the simulation of fire spread onboard naval vessels[J]. Fire Safety Journal, 2017,91:441-450.
[17] Bringmann K, Keusch R, Lengler J, et al. Greedy routing and the algorithmic small-world phenomenon[C]// Proceedings of the ACM Symposium on Principles of Distributed Computing. 2017:371-380.
[18] 赵明. 黄金分割法[J]. 科学与文化, 1997(5):55.
[19] 张松海,施心陵,李鹏,等. 多峰函数优化的黄金分割斐波那契树优化算法[J]. 电子学报, 2017,45(4):791-798.
[20] 刘长平,叶春明. 具有 Lévy 飞行特征的蝙蝠算法[J]. 智能系统学报, 2013,16(3):240-246.

(上接第 27 页)

[38] Corro L D, Gemulla R. ClausIE: Clause-based open information extraction[C]// Proceedings of the 22nd International Conference on World Wide Web. 2013: 355-366.
[39] Melamud O, Berant J, Dagan J, et al. A two level model for context sensitive inference rules[C]// Meeting of the Association for Computational Linguistics. 2014: 1331-1340.
[40] Han Xianpei, Sun Le. Context-sensitive inference rule discovery: A graph-based method[C]// Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. 2016:2902-2911.
[41] Banko M, Cafarella M J, Soderland S, et al. Open information extraction from the Web[C]// Proceedings of IJCAI. 2007:2670-2676.
[42] Fei Wu, Weld D S. Autonomously semantifying Wikipedia[C]// Proceedings of the 16th ACM Conference on Information and Knowledge. 2007:41-50.
[43] 秦兵,刘安安,刘挺. 无指导的中文开放式实体关系抽取[J]. 计算机研究与发展, 2015,52(5):1029-1035.
[44] 郭喜跃. 面向开放领域文本的实体关系抽取[D]. 武汉:华中师范大学, 2016.
[45] 李颖,郝晓燕,王勇. 中文开放式多元实体关系抽取[J]. 计算机科学, 2017,44(S1):80-83.
[46] Pasca M. Organizing and searching the World Wide Web of facts-step two: Harnessing the wisdom of the crowds[C]// Proceedings of the 16th International Conference on World Wide Web. 2007:101-110.
[47] Socher R, Huval B, Manning C D. Semantic compositionality through recursive matrix-vector spaces[C]// Proceedings of 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012:1201-1211.
[48] Zeng Daojian, Liu Kang, Lai Siwei, et al. Relation classification via convolutional deep neural network[C]// Proceedings of International Conference on Computational Linguistics. 2014:2335-2344.
[49] Miwa M, Bansa M. End-to-end relation extraction using LSTMs on sequences and tree structures[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016:1105-1116.
[50] Lin Yankai, Liu Zhiyua, Sun Maosong. Neural relation extraction with multi-lingual attention[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017:34-43.
[51] Lin Yankai, Shen Shiqi, Liu Zhiyuan, et al. Neural relation extraction with selective attention over instances[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016:2124-2133.
[52] 孙建东,顾秀森,李彦,等. 基于 COAE2016 数据集的中文实体关系抽取算法研究[J]. 山东大学学报(理学版), 2017,52(9):7-12.
[53] 林衍凯,刘知远. 基于深度学习的关系抽取[EB/OL]. http://www.cipsc.org.cn/qngw/?p=890, 2016-09-14.
[54] 刘绍毓. 实体关系抽取关键技术研究[D]. 郑州:解放军信息工程大学, 2015.