实体关系的弱监督学习抽取方法

王 政[1] 朱礼军[1] 徐 硕[2]
（1. 中国科学技术信息研究所，北京 100038；
2. 北京工业大学经济与管理学院，北京现代制造业发展研究基地，北京 100124）

摘要：许多相关综述工作往往侧重于通用领域的实体关系抽取，没有充分体现弱监督学习实体关系抽取的重要性。文章立足于科技情报领域，在综述大量文献的基础上，剖析半监督、远程监督和无监督等3种弱监督实体关系抽取方法的逻辑关系，并总结了各自的优缺点。基于弱监督学习的实体关系抽取方法可以部分解决标注数量不足的问题。特别是，针对不同的科技情报应用，多种弱监督实体关系抽取方法的综合运用可以取得显著效果，但是精确度偏低的问题在未来相当长的时间内仍然是弱监督学习实体关系抽取的主要挑战。

关键词：弱监督学习；实体关系抽取；半监督学习；远程监督学习；无监督学习
中图分类号：G35 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2018.02.017

Review of Weakly Supervised Relation Extraction

WANG Zheng[1], ZHU Lijun[1], XU Shuo[2]
(1. Institute of Scientific and Technical Information of China, Beijing 100038; 2. Research Base of Beijing Modern Manufacturing Industry Development, School of Economics and Management, Beijing University of Technology, Beijing 100124)

Abstract: Although recent related reviews focused on general knowledge, they were unaware of the importance of weakly supervised relation extraction. So, on the basis of science and technology information domain, this work analyzed the logical relations among semi-supervised, distant supervised and unsupervised methods by reviewing, and the resulting advantages and disadvantages were summarized. It's believed that weakly supervised learning could partially solve the scarcity of labeled data. Especially for specific technology information applications, integrated weakly supervised methods may make giant improvement. But in the near future, the main challenge for weakly supervised relation extraction is still the problem of precision.

Keywords: weakly supervised learning, relation extraction, semi-supervised learning, distant supervised learning, unsupervised learning

0 引言

大数据使得许多利用传统方法难以解决的问题变得可行。例如，在医疗问答系统中如果知道“马钱子”和“肾毒性”成“正相关”的关系，那么问题“低蛋白血症应该吃什么药?”对应的答案中就可以筛除含有马钱子的中药药方。但是，表达“马钱子”与“肾毒性”关系的语句往

作者简介：王政（1992—），男，中国科学技术信息研究所硕士研究生，研究方向：知识图谱构建、主题模型；**朱礼军**（1973—），男，博士，中国科学技术信息研究所研究员，研究方向：Semantic Web、Web Service和知识技术在科技信息服务中的应用；**徐硕**（1979—），男，博士，北京工业大学经济与管理学院校聘教授，研究方向：数据挖掘、机器学习和知识工程等（通讯作者）。

基金项目：国家自然科学基金项目“基于论文和专利资源的技术机会发现研究”（71403255）。
收稿日期：2017 年 7 月 28 日。

往存在于专业网站、学术文献和科技类图书等科技文献资源中，因此基于科技文献资源的关系抽取为此类问题的解决带来了希望。

早在1996年，由美国军方背景支持的MTU会议就意识到了这一点，提出要通过多种手段提升人类的数据利用能力，并对这一目标进行了具体而详细的阐述[1]。实体关系抽取在其中起到了承上启下的作用，其准确率和效率直接影响后续任务（如事件抽取、情感分析等）的性能，因此备受国内外研究者们的重视[2-4]。

近年来，许多学术或者商业项目在通用领域开展了大量的关系抽取实践，形成了YAGO2[5]、NELL[6]、Freebase[7]、DBpedia[8]、Google Knowledge Vault[9]等知识库。在结构上，这些知识库中主要包含了大量的二元关系，如Person-Org关系、Org-Address关系等；偶尔也存在一些多元关系（N-ary Relation），如“A在B和C中间”[10]，但并不占主流。从构建方法上来说，为了从大量无结构或者半结构的语料中构建知识库，主要应用监督方法、远程监督方法、半监督方法和无监督方法。

对于科技情报领域，监督实体关系抽取方法不具有优势。因为监督实体关系抽取器的训练需要首先通过全面、高质量的标注数据训练实体关系抽取器，然后再通过实体关系抽取器从未标注数据中抽取实体关系。以常用的ACE（Automatic Content Extraction）语料为例，其中包含了超过1000个文档，每个文档中的实体对被标注了5～7个主要关系与23～24个次要关系，共计16771个关系实例。然而，科技情报往往涉及多个领域，专业性强、标注成本高、含有大量专有名词、关系类型不固定。为了达到通用领域实体关系抽取的类似水平，需要投入大量的人力、物力和财力资源。

弱监督学习方法，即半监督学习、远程监督学习和无监督学习，则可有效解决这一问题：无论标注数据中是否存在错误、带有噪音，还是标注数据原本不是用于意向目标，抑或只存在一些先验知识、根本没有标注数据。上述方法均可以用于实体关系抽取。特别是，近年来，随着实体关系抽取研究的深入，这3种方法常常相互启发、互相配合，在同一套项目中作为一个整体出现[11-13]。

尽管弱监督学习实体关系抽取前景乐观，但是相关综述性文献比较少。如Konstantinova[2]的综述重点在于对通用语料的实体关系抽取进行一个整体性的阐述，客观上缺乏对科技情报的适用性。而其他学者如Bach和Badaskar[3]、车万翔等[4]所做的综述，由于历史原因仅限于监督实体关系抽取方法。为了促进弱监督实体关系抽取在科技情报界的应用，本文拟按照对标注数据的要求，对弱监督学习的发展历程及其半监督、远程监督和弱监督学习3种方法进行描述和分析。

1　弱监督学习抽取方法的发展历程

随着信息技术的发展，互联网上所承载的资源日益增加，利用方式不断丰富。而要对这些无结构或半结构的信息资源进行深入挖掘与利用，需要将它们进行结构化。而从无结构、半结构数据构建结构化数据的方法之一，就是实体关系抽取。如图1所示。MUC[1]会议认为，实体关系抽取任务是未来发展的一个重要方向，并首先进行了定义。传统上，研究者们往往使用监督学习方法将实体关系抽取视作分类问题，通过以核函数[14]为代表方法从标注数据中学习关系抽取器。尽管该方法取得了不小的进展，但面对越来越多的数据与不同领域的实体关系抽取需求，其数据标注成本越来越高。

1998年，谷歌利用PageRank等算法在信息检索方面进行了成功的尝试，人们只需要输入关键词即可得到相关信息。但是，在没有更自然、更精准的检索服务的情况下，用户仍然需要翻阅多个页面才能获得自己想要的结果。而提供更自然、更精准的检索服务，显然需要进行实体关系抽取。

同年，Brin[15]使用半监督学习做出的工作引发了研究者们的注意：他使用少量数据作为“种子”，对“作者—书籍”关系进行抽取。他从

“种子”中获得能够匹配关系的模板，进而可以匹配新的关系实例。虽然这种方法受限于专业领域知识背景和“种子”的质量，但是它证明，减少数据标注依赖是有可能的。

随着Web 2.0为基础的多种互联网服务的发展，维基百科等公共知识库吸引了越来越多的目光。因此，一种可行的思路是通过这些公共知识库拓展标注数据的来源，利用知识库中半结构化的数据为结构化数据提供帮助，这种方法被称作远程监督学习方法。很多基于维基百科的结构化知识库的发展，如Freebase[7]、DBpedia[8]等，为远程监督学习奠定了应用基础。

然而，许多具有专业知识背景的实体关系抽取项目仍然无法找到合适的知识库支持。对于这种情况，2008年，谷歌提出了OpenIE方法。该方法通过无监督学习实体关系抽取彻底摆脱了标注数据的限制，更加适用于多领域、大规模数据。实践表明，无监督学习实体关系抽取方法极大地改善了谷歌的检索质量，使用者可以通过更自然的方式获得更精准的实体关系抽取结果。

至此，上述3种方法形成了与监督学习方法截然不同的实体关系抽取思路，即弱监督学习实体关系抽取。在之后的实体关系抽取发展过程中，很多实体关系抽取模型都会综合利用这3种方法，以全面测试模型的性能。因此，本文对3种方法进行综述，以帮助读者全面了解弱监督学习实体关系抽取。

2　半监督学习

半监督学习已经成为弱监督学习实体关系抽取中应用最广泛的方法，其标志性的自训练[15-16]过程如图2所示。

（1）从一个较小的数据集开始，标注出其中的关系实例，这些关系实例被称作“种子”。

（2）从“种子”中提取模板。

（3）通过模板在非“种子”语料中提取新的

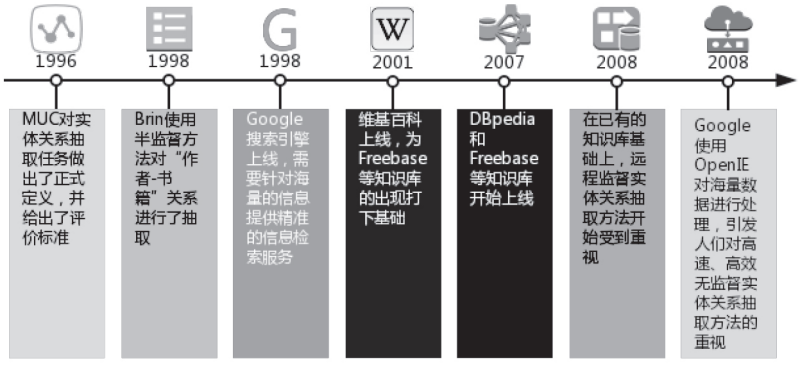

图1　弱监督学习发展历程中的关键节点

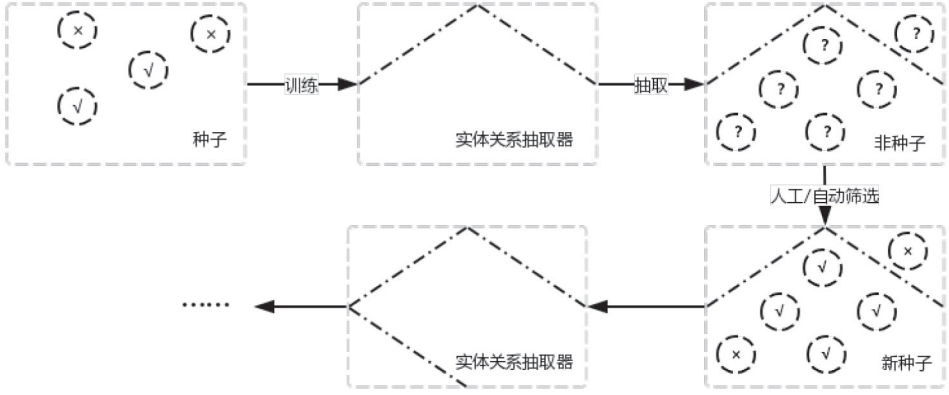

图2　半监督学习训练过程

实体关系实例，并将这些实例作为新的种子。

（4）从步骤二开始执行，直到循环终止条件达成。

其目标是通过很少的标注数据训练出较好的实体关系抽取模型，并抽取出大量的关系实例。例如要从互联网上抽取“书—作者”关系，Brin[15]只使用了5个关系实例作为种子，就可以从自然语言文本、URL、超链接中为当时尚不完善的文献数据库补充15257个实例。类似的关系还包括“科研机构—作者”、作者合著、机构合作、母体文献、项目来源等[17]。

但是，少量的人工标注数据容易产生语义漂移，误导实体关系抽取模型学习到不合适的“种子”和模板。解决这个问题的基本思路是加强人的监督。比如利用模板与关系实例的对偶性[15]将模板视作对实例的抽象，将实例视作模板所表示关系的具体实现。Brin选择了一种字符串匹配模板，既方便在计算机上实现，也方便研究人员的阅读与理解，从而可以把错误的模板和匹配的错误实例去掉，在保留346个模板的情况下抽取到大量实例。

这种方法的缺点是：有时候要抽取的实体关系太多，人工筛选仍然耗时耗力。因此，在上述半监督学习自训练过程的基础上，Blum和Mitchell[18]通过协同训练改进了上述自训练过程的后三步，即：

（2）用每个关系的“种子”训练对应的实体关系抽取器。

（3）通过实体关系抽取器对非“种子”语料提取新的实体关系实例。

（4）对新抽取出来的实体关系实例进行筛选，得到新的“种子”。

很明显，第三步可以利用不同关系之间的相互作用，通过人工编写的规则筛除不合适的实例。但这基于对抽取关系足够精细的认知，筛选规则的编写事实上受到研究人员认知的限制，因为很难区分什么是“特例”，什么是“错误”。例如《黑客帝国》的导演沃卓斯基兄弟实体对，因为兄长做了变性手术，所以有的人认为“姐弟”关系在特定的时间也成立。这种加上时间、地点等条件的关系也被称作“事件”[19]。

另一种思路被称作“避免密集区域改变”[20]：如果一个实例和其他实例相似度较低，那么这个实例有可能是错误的；如果一个实例和其他实例相似度较高，那么其错误的可能性就较低。反过来，如果有多种关系可能出现于某个实体对时，那么相似的关系更可能同时出现，相似度较低的关系则要进行适当的割舍。因此，如果“协同训练”利用的是关系之间的“协同”性判断关系实例是“特例”还是“错误”，那么这种“协同性”同样可以作用于数据之间：将非“种子”语料分割成若干份，分别训练实体关系抽取器，此抽取器判断为某关系的实例可能被其他抽取器判断为非实例，这样的实例因此可以被筛除。

总之，半监督学习在“种子”筛选方面还有很长的路要走，目前看来有两个发展方向：一是提高模型训练速度；二是将“种子”的筛选方法与对目标关系的描述结合起来，特别是结合逻辑描述与概率描述两种手段。

3　远程监督学习

远程监督的目标则是尽可能增加标注数据，其具体方法是将某些结构化的数据源转化为可用的标注数据集。这样的数据集通常以各种人工构建的知识库形式呈现，如Kozareva等[21]研究了如何利用维基百科发现实体关系。在这样的知识库基础上，可以总结远程监督具有以下一般流程。

（1）从现知识库中收集关系实例，如Craven和Kumlien从人工构建的生物学Yeast Protein Database知识库中收集了1213个“亚细胞定位”关系实例。

（2）将关系实例中的实体对分离出来，即“亚细胞定位”关系对应的蛋白质实体和“亚细胞位置”实体组成的实体对。

（3）从待处理语料中根据不同规则找到对应关系的实例。

（4）使用上述标注数据训练实体关系抽取

器。

该流程的重点是第二步和第三步，即如何收集实体对并将知识库中对应的关系映射到无结构文本中。针对不同资源可以采取不同的措施，Kozareva等[21]在第二步首先使用维基百科词条间的超链接建立图结构，在这个结构中，如果“度”满足一定条件，即可认为这两个实体具有一定关系。如Craven和Kumlien[22]认为一个句子只要同时包含蛋白质实体和“亚细胞位置”实体，即可将对应的实体对标注为“亚细胞定位”实体关系。

虽然Craven和Kumlien[22]的方法简单有效，能够从633个句子中收集到336个关系实例。但是其假设过强，每一个同时包含两个实体的句子都会表述这两个实体在知识库中的对应关系[23]，这可能导致如图3所示的各种问题。例如，一个句子中如果出现“乔布斯”和“苹果公司”这两个实体，这个句子很可能表述了“CEO−of”关系。但是在知识库中这两个实体往往还构成“Founder−of”关系，如何判断某一句话到底要表达哪种关系就出现问题了。

这个问题的解决方案是将一种关系看作另一种关系的“噪音”。“沃兹尼亚克”与“苹果公司”构成“Founder−of”关系而不构成“CEO−of”关系，因此可以用确定为“Founder−of”的关系实例来生成实体关系抽取器，然后判断某句话中“乔布斯”与“苹果公司”是否构成“Founder−of”关系。根据这种想法，Yao等[11]通过远程监督方法将Mintz等[12]获得的关系实例作为观测得到的先验知识加入主题模型并进行了聚类。如果先验中一个实体对被标注了两种关系，接下来的聚类过程自会判断这两种关系是否成立。

不难发现，在其他研究中，实体关系抽取的目标是根据语料给出的特征判断实体对具体表现为什么关系。而在远程监督中，目标变成了根据实体对的已知关系对包含这个实体对的语料特征的表述进行判断。

这种视角变换引起了Surdeanu等[13]的注意，他们提出了MIML（Multi−instance Multi−Label）模型以允许某个关系实例表述多种关系。特别是在知识库相当全面的情况下，如果某个实体对存在多种关系，这种假设显然更具有普适性和实用性：如果一个非常全面的知识库中某个实体对不表述某种关系，那么对应的关系实例也应当斟酌是否表述该关系。从更高的层面来说，“多种关系在实体对层面上存在共现”，这样的逻辑关系比Yao等[11]的“多种关系在文档层面存在共现”更有说服力，这为结合使用半监督和远程监督方法提供了途径。

4　无监督学习

维基百科“中国”词条的信息框（InfoBox）中，“北京”与“中国”的关系是“首都”（Capital）。通过这样一个关系实例，我们可以提取相应的特征，包括其在信息框的HTML代码

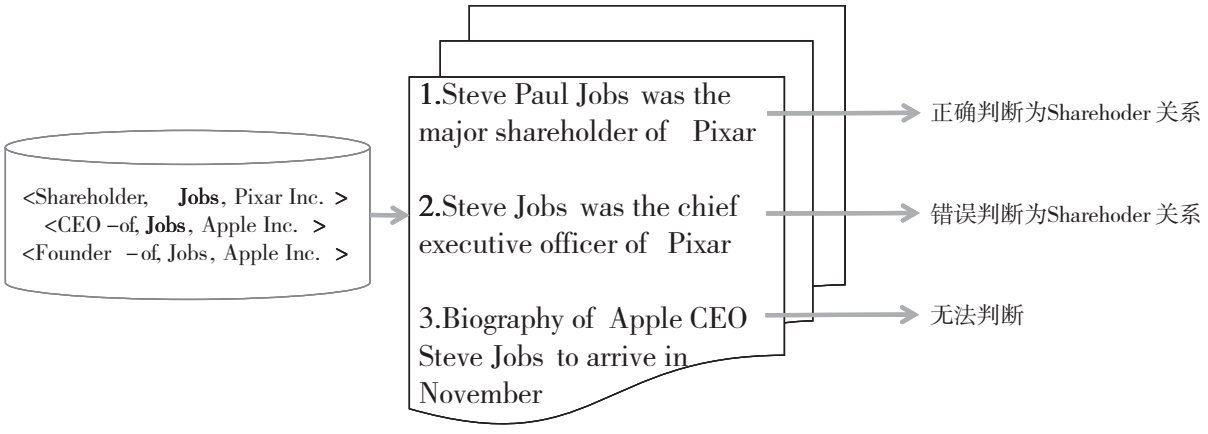

图3　远程监督实体关系抽取可能遇到的各种情况

中所处的相对位置，“首都”这个词以及对应的自然语言特征等。一般认为，这些特征适用的范围不仅限于关系实例，也适用于关系本身的其他实例，这被称作“平移不变性”[24]。仍然以维基百科为例：中国和美国词条中都出现了“最大城市”的关系实例，显而易见，这种实体关系的发现并不需要任何监督（图4）。

为了发现这种“平移不变性”，OpenIE等[25]设计了8个领域知识无关的词法—句法模板用以匹配相关特征。研究者认为，这些模板能够匹配95%以上的实体关系实例，并为实体关系的判断提供足以判断具体关系的特征，Nguyen等[26]则通过另外训练的CRF模型识别特征所对应的关系。这种方式简单、有效、适合并行化，在理想的情况下只要数据足够多，总能抽取到所有正确的实体关系实例。

其缺点是抽取出来的关系实例有13%“碎片化”，有7%“无信息”[27]。如“The guide contains dead links and omits sites.”和“gave birth to”，按照OpenIE的模板可能抽取出“contain omit”关系和“give”关系。对此，Nguyen等的解决方案是通过观察语料中关系实例的具体形式，加入新的词法和句法约束形成新的模板，将原来省略掉的实体关系标注成本转移到了模板设计方面。虽然由于OpenIE对关系基本上不进行聚类，所以它不会把不同的关系错误判断为一类，但这同样导致缺少对特征的归纳总结过程。

因此，使用无监督学习的研究者仍然需要一些可用的先验知识来实现关系本身的消歧。在先验知识的帮助下结合Yao等[11]的Rel−LDA和Type−LDA模型，以模型训练速度与实体关系抽取速度为代价，获得相当高的无监督学习实体关系抽取精确度，不论这种知识是远程监督提供的还是监督学习语料提供的。值得一提的是，先验知识导入时，在OpenIE中先验知识以模板的形式存在，情报科学语料模板的编写需要专家的经验与专业知识，而Rel−LDA和Type−LDA完全不需要这一点，它们会自行从先验知识中学习关系对应的统计学特征。

5　结语与讨论

如表1所示，弱监督学习实体关系抽取主要解决了监督学习对标注数据的需求问题，这对于科技信息（情报）服务业的检索引擎、垂直问答系统[28−30]以及面向专业领域的机器翻译[31−32]等有极为重要的意义。而针对不同的应用目标，3

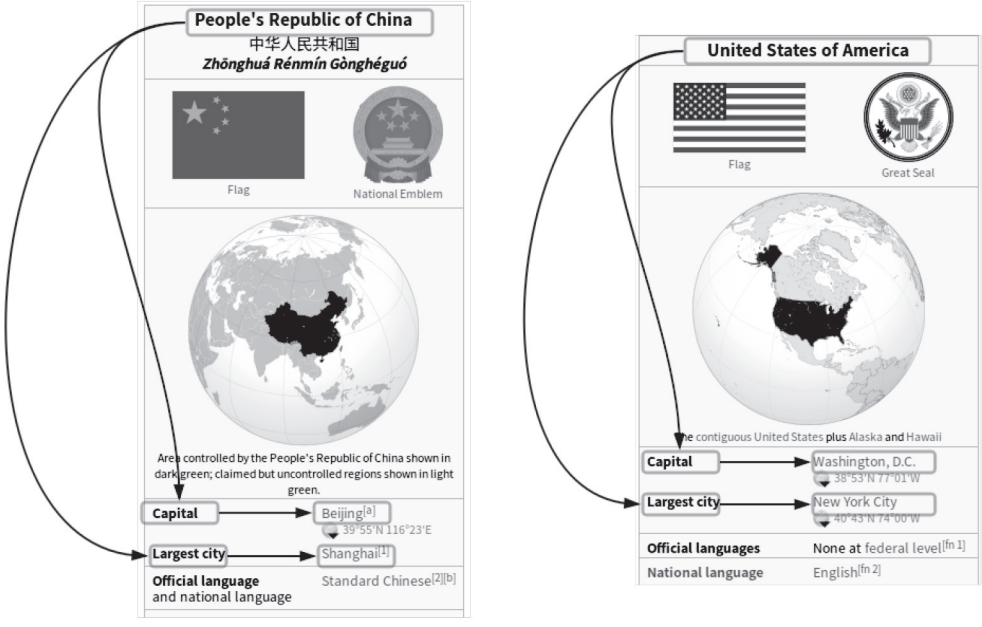

图4　Capital-of关系抽取中的平移不变性

表 1　弱监督学习实体关系抽取 3 种方法的一般特点

	半监督	远程监督	无监督
标注需求	小批量“种子”	任何可用的结构化数据	无
规模数据可拓展性	较小	一般	强
速度	与“种子”筛选方式相关	一般	快
主要问题	语义漂移	知识库噪音	关系消歧

种弱监督方法因其不同特点有不同的适用范围。

从对标注数据的需求看：半监督学习可以根据已标注的少量“种子”从未标注数据中学习得到目标关系实例，部分解决标注数量不足的问题；远程监督学习可以使用原本不是用于意向目标的知识库扩展实体关系抽取器训练数据来源；在无训练语料的情况下，无监督学习可以利用关系实例之间的“平移不变性”进行关系抽取，这在语料数量比较多的情况下可行性较强。

从适用数据的规模看：半监督学习方法在缺少合适“种子”和筛选方式的情况下，较容易出现语义漂移，因此应用于大规模数据有困难；在高质量、大规模知识库的支持下，远程监督学习可以应用于一般规模的数据；无监督学习由于没有标注数据的制约，只要模型设计合理即可在大规模数据的基础上进行实体关系抽取。

从弱监督学习实体关系抽取的主要短板上看：半监督学习受限于“种子”，容易产生语义漂移问题；远程监督无法避免数据库带来的噪音；而无监督学习在同一关系的不同表述上，消歧能力有待加强。这些问题可以总结为精度不高，这在数量较小的专业领域的语料上尤其严重。

尽管不同的弱监督实体关系抽取方法有不同的特点，但多种方法互相借鉴才是未来发展的主流方向。在一段时间内，科技情报领域实体关系抽取需要综合使用远程监督拓展来自专业领域的知识特征，结合待抽取关系的一般特点专门构建模型，并且选用有代表性的数据作为先验知识，这样才能在较少的标注数据上达到较好的实体关系抽取效果。

参考文献

[1] GRISHMAN R, SUNDHEIM B. Message understanding conference−6: a brief history[C]//proceedings of the 16th conference on computational linguistics, 1996: 466−471.

[2] KONSTANTINOVA N. Review of relation extraction methods: what is new out there?[J]. Communications in Computer & Information Science, 2014, 436(1):15−28.

[3] BACH N, BADASKAR S. A review of relation extraction [R]. Carnegie Mellon University, 2007.

[4] 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.

[5] HOFFART J, SUCHANEK F M, BERBERICH K, et al. YAGO2: Exploring and querying world knowledge in time, space, context, and many languages[C]//proceedings of the 20th international conference companion on world wide web, 2011: 229−232. DOI: 10.1145/1963192.1963296.

[6] MITCHELL T, COHEN W, HRUSCHKA E, et al. Never−ending learning[C]//proceedings of the 29th AAAI conference on artificial intelligence, 2015: 2302−2310.

[7] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//proceedings of the 2008 ACM SIGMOD international conference on management of data, 2008: 1247−1250. DOI: 10.1145/1376616.1376746

[8] AUER S, BIZER C, KOBILAROV G, et al. DBpedia: a nucleus for a web of open data[J]. Lecture Notes in Computer Science, 2007, 4825: 722−735. DOI: 10.1007/ 978−3−540−76298−0_52.

[9] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a web−scale approach to probabilistic knowledge fusion[C]//proceedings of the 20th ACM SIGKDD international conference, 2014: 601−610. DOI: 10.1145/2623330.2623623.

[10] GRIM P, BARWISE J, ETCHEMENDY J, et al. Language, proof and logic[M]. [S.l.]: Center for the Study of Language and Inf Publications, 2001,7(3):19−20.

[11] YAO L, HAGHIGHI A, RIEDEL S, et al. Structured

relation discovery using generative models[C]//proceedings of the 2011 conference on empirical methods in natural language processing, 2011: 1456−1466.

[12] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]// proceedings of the 47th annual meeting of the association for computational linguistics, 2009: 1003−1011. DOI: 10.3115/1690219.1690287.

[13] SURDEANU M, TIBSHIRANI J, NALLAPATI R, et al. Multi−instance multi−label learning for relation extraction[C]//proceedings of the 2012 joint conference on empirical methods in natural language, 2012: 455−465.

[14] ZELENKO D, AONE C, RICHARDELLA A , et al. Kernel methods for relation extraction[J]. Journal of Machine Learning Research, 2003(3): 1083−1106.

[15] BRIN S. Extracting patterns and relations from the world wide web[C]//international workshop of the world wide web and databases, 1998: 172−183. DOI: 10.1007/10704656_11.

[16] ZHU X. Semi−supervised learning literature survey[R]. Computer Sciences, University of Wisconsin−Madison, 2008. DOI: 10.2200/S00196ED1V01Y200906AIM006.

[17] 张晗, 徐硕, 乔晓东. 融合科技文献内外部特征的主题模型发展综述[J].情报学报, 2014(10): 1108−1120.

[18] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co−training[C]//proceedings of the 11th annual conference on computational learning theory, 1998: 92−100. DOI: 10.1145/279943.279962.

[19] 赵妍妍, 秦兵, 车万翔, 等. 中文事件抽取技术研究[J]. 中文信息学报, 2008, 22(1): 3−8.

[20] SEEGER M. Learning with labeled and unlabeled data[C]//The European symposium on Artificial neural networks, 2002: 1−62. DOI: 10.1109/IJCNN.2002.1007592.

[21] KOZAREVA Z, RILOFF E, HOVY E. Semantic class learning from the web with hyponym pattern linkage graphs[C]//proceedings of the 46th annual meeting of the association for computational linguistics, 2008(June): 1048−1056.

[22] CRAVEN M, KUMLIEN J. Constructing biological knowledge bases by extracting information from text sources[C]//proceedings of the international conference on intelligent systems for molecular biology, 1999: 77−86.

[23] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]// proceedings of the 47th annual meeting of the association for computational linguistics, 2009, 2: 1003−1011.

[24] BORDES A, USUNIER N, WESTON J, et al. Translating embeddings for modeling multi−relational data[C]// advances in NIPS, 2013, 26: 2787−2795. DOI: 10.1007/s13398−014−0173−7.2.

[25] ETZIONI O, BANKO M, SODERLAND S, et al. Open information extraction from the web[J]. Communications of the ACM, 2008, 51(12): 68. DOI: 10.1145/1409360.1409378.

[26] NGUYEN N T H, MIWA M, TSURUOKA Y, et al. Open information extraction from biomedical literature using predicate−argument structure patterns[C]//the 5th international symposium on languages in biology and medicine, 2013: 51−55.

[27] FADER A, SODERLAND S, ETZIONI O. Identifying relations for open information extraction[C] //proceedings of the 2011 conference on empirical methods in natural language processing, 2011: 1535−1545. DOI: 10.1234/12345678.

[28] 刘杰, 樊孝忠, 王涛. 基于本体的受限领域问答系统研究[J]. 广西师范大学学报(自然科学版), 2009, 27(1): 169−172.

[29] YIH W T, CHANG M W, HE X, et al. Semantic parsing via staged query graph generation: question answering with knowledge base[C]//proceedings of the 53rd annual meeting of the association for computational linguistics, 2015: 1321−1331.

[30] LIJUN Z, Ning Z. Research on natural language question analysis based on knowledge organization system[D]. Beijing: Institute of Scientific and Technical Information of China, 2016.

[31] 达瓦·伊德木草, 艾山·吾买尔. 实例统计翻译混合策略的汉民病历翻译的研究[J]. 新疆大学学报(自然科学版), 2015(1): 68−73.

[32] LAO N, SHIMA H, MITAMURA T, et al. Query expansion and machine translation for robust cross−lingual information retrieval[C]//proceedings of the 7th NTCIR workshop meeting on evaluation of information access technologies, 2008: 140−147.