基于卷积神经网络的中文医疗弱监督关系抽取

刘 凯 符海东 邹玉薇 顾进广
(武汉科技大学计算机科学与技术学院 武汉 430065)
(智能信息处理与实时工业系统湖北省重点实验室 武汉 430065)

摘 要 随着医疗领域受到越来越多的关注,自然语言处理的理论和应用逐渐拓展到该领域,其中信息抽取技术在该领域的应用成为研究热点。针对信息抽取技术在医疗领域实体关系抽取中的应用,提出一种基于卷积神经网络的弱监督关系抽取方法。该方法通过添加人工规则使训练语料带有实体关系标签,然后将该弱关系训练语料转换为向量特征矩阵,并输入到卷积神经网络进行分类模型训练,最终实现实体关系抽取。实验结果表明,该方法比常规机器学习方法更加准确高效。

关键词 自然语言处理,实体关系抽取,弱监督,卷积神经网络
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.10.045

Chinese Medical Weak Supervised Relation Extraction Based on Convolution Neural Network

LIU Kai FU Hai-dong ZOU Yu-wei GU Jin-guang
(School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China)
(Hubei Province Key Laboratory of Intelligent Information Processing and Real Time Industrial System, Wuhan 430065, China)

Abstract With medical field are receiving more and more attention, the theory and application of natural language processing began to expand the field, and information extraction technology in the field of application has become a research hotspot. In this paper, based on the application of information extraction technology in medical domain entity relation extraction, a weak supervised relation extraction method based on convolution neural network was proposed. This method adds the artificial rules to the training corpus with the entity relation label, and then transforms the weak relation training corpus into the vector characteristic matrix, next inputs it into the convolution neural network for training the classification model, and finally realizes the entity relation extraction. The experimental results show that the method is more accurate and efficient than the conventional machine learning method.

Keywords Natural language processing, Entity relation extraction, Weak supervision, Convolutional neural network

1 引言

目前人们浏览到的资讯大多以非结构化的形式存储在互联网上,其中蕴藏着丰富的知识,包含了大量的实体关系实例,因此面向中文的实体关系抽取成为了自然语言处理的重要工作之一。但随着中文医疗领域的数据激增,常规的面向中文的实体关系抽取方法在该领域的适用性与准确性开始下降,因此需要基于传统关系抽取方法在中文医疗领域进行细化改进。

实体关系是构建知识库的关键,近年来,研究人员提出了各种方法来实现面向非结构化文本的实体关系抽取[1],根据抽取方法对训练语料的不同要求可以分为:无监督关系抽取、有监督关系抽取、弱监督关系抽取、开放关系抽取4种范式。无监督关系抽取[2]不需要过多的人为干预,该方法可以基于相似模板聚类来发现新的关系,但其缺点是得到的关系不具有语义化信息,难以规则化,如果要进一步构建知识库还需要人工操作。开放关系抽取[3]是为了处理大量异构网络数据而设计的,其所抽取的关系类型是不受限制的,数量也不固定,对于特定领域的实体关系抽取的效率不高。有监督学习方法的性能最好,目前占据主导地位。但是在面对海量的中文医疗数据时,现有的有监督学习方法需要大量人工标注的训练语料,而人工标注语料费时费力,一致性较差。为解决该问题,本文提出一种弱监督学习方法来实现训练语料的半自动化生成。该方法是一种基于噪声训练数据的半监督方法,利用知识库中已有的关系实体对,通过弱监督规则的添加自动获得训练语料,有效解决了训练语料不足的问题。有监督学

到稿日期:2016-08-05 返修日期:2016-11-21 本文受湖北省自然科学基金(2013CFB334)资助。
刘 凯(1992-),男,硕士,主要研究方向为语义网信息抽取,E-mail:2507161048@qq.com;符海东(1971-),男,博士,主要研究方向为语义网信息抽取与智能化语义表达;邹玉薇(1991-),女,硕士,主要研究方向为语义网络实体关系抽取与知识图谱构建;顾进广(1974-),男,博士,主要研究方向为分布式计算、智能信息处理、语义Web及软件工程,E-mail:simon@wust.edu.cn(通信作者)。

习方法往往将关系抽取问题转化为多分类问题，采用常用的机器学习分类器来实现关系抽取。使用常用分类器时不可避免地需要提取特征[21]，传统的特征提取方法一般采用词法分析、句法分析等基本的自然语言处理工具，其主要有两个缺点：1)特征提取完全凭经验进行，提取特征质量依赖于现有的自然语言处理工具的准确率，存在误差累积问题；2)某些语言缺少自然语言处理所需的资源，不利于传统的特征提取方法的推广。为解决上述问题，本文使用卷积神经网络作为实体关系抽取的分类器，通过弱监督加入的实体关系标注自动提取特征，然后将提取的特征进行特征向量矩阵化并输入到卷积神经网络中，以完成模型训练，最终实现实体关系抽取。

本文第 2 节主要介绍已有的面向中文的关系抽取方面的相关研究；第 3 节详细介绍基于卷积神经网络的面向中文医疗领域实体关系抽取方法；第 4 节通过实验验证了本文方法的有效性，并通过与传统方法的对比说明了该方法在中文医疗领域的实用性与高效性；最后总结全文。

2 相关研究

关系抽取是信息抽取的任务之一，根据参与概念的多少可以将其分为二元关系抽取和多元关系抽取[4]。近年来，较多实体抽取方法被提出，目前占据主导地位的是基于机器学习的有监督学习方法和弱监督学习方法。有监督学习方法根据训练样本表示方法的不同可以分为基于特征向量的方法和基于核函数的方法。基于特征向量的方法着重提取有区分度的特征来描述关系实例中的局部特征或实体特征。文献[5]中的方法综合考虑实体本身、实体类型、依存树和解析树等特征，使用最大熵分类器判断实体间的关系；Zhou 等[6]系统地研究了如何把包括基本词组块在内的各种特征组合起来，探讨了各种语言特征对关系抽取性能的贡献；Jiang 等[7]通过统一的特征空间表达形式来研究不同特征对关系抽取性能的影响，其中特征空间可划分为序列、句法树和依存树等特征子空间。基于特征向量的方法尽管速度快且较为有效，但其缺点是在转换结构化特征时需要显式地给出一个特征集合，由于实体间语义关系表达的复杂性和可变性，要进一步提高关系抽取的性能较为困难，因为很难再找出适合语义关系抽取的有效的新词汇、句法或语义特征。不同于特征向量的方法，基于核函数的方法直接以结构树为处理对象，在计算关系之间的距离时直接在高维的特征空间中使用核函数隐式地计算对象之间的距离，不用枚举所有的特征也可以计算向量的点积，这表明实体关系更加灵活。基于核函数的关系抽取最早由 Zelenk 等[8]提出，他们在文本的浅层句法树的基础上定义了树核函数，并设计了一个计算树核函数相似度的动态规划算法，然后通过支持向量机(SVM)和表决感知器(VotedPerceptron)等分类算法来抽取实体间的语义关系。Zhang 等[9]融合卷积树核函数(Convolution Tree Kernels，CTK)和线性核函数，综合考虑了影响实体间语义关系的平面特征和结构化特征，利用卷积树核函数来计算包含实体对的句法树之间的相似度，使用线性核函数计来算实体属性(如实体类型等)间的相似度。Zhou 等[10]提出了最短路径包含树核，将语义关系实例表示为上下文相关的最短路径包含树，并采用上下文相关的核函数计算方法将该核函数同基于特征的方法相结合。

基于弱监督学习的关系抽取最早由文献[11]提出，被用于从学术文献的摘要中抽取蛋白质与基因之间的关系。文献[12]将弱监督关系抽取看作一个多示例问题，假设只要求在回标出来的所有句子中至少存在一个句子能表示两个实体间的关系。其将所有回标的句子看作一个包，其中每一个句子就是包中的一个示例，从而解决回标噪音的问题。文献[13]不但对噪声训练数据进行建模，并且对实体对可能属于多个关系类型的问题进行建模，提出了基于概率图模型的多实例多标签模型，在以 Freebase 为知识库和纽约时报作为回标语料的数据上进行实验，结果表明其模型提升了原始方法的抽取效果。文献[14]利用信息检索中的伪相关反馈来解决样本噪音问题。文献[15]借助样本空间的密度信息，求出密度中心店来准确地反映数据的空间几何特征，在此基础上建图，利用标记传递方法使得相似的顶点尽可能被赋予相同的类别标记，结合大量未知模式样本进行弱监督学习。

卷积神经网络是人工神经网络的一种，过去多用于图像识别领域。当图像与文本直接作为网络的输入时，卷积神经网络可以直接从原始数据中构造大量的冗余特征，减少了传统方法中复杂的特征提取与数据重建过程，当数量足够大时，网络也能自动选出有效信息，因此，近年来卷积神经网络也被有效地应用于各种自然语言处理任务中，包括词性标注、名词短语识别、命名实体识别、语义角色标注[16]、语义解析[17]、句子模型和分类[18]、关系分类和抽取[19]。

3 中文医疗弱监督关系抽取方法

弱监督学习关系抽取依赖于结构化的知识库。现有的医疗知识是海量分布在分离式的网络媒介中的，其中最普遍的实体及实体关系载体是网页。关系抽取的前提是从这些分离式的网页中抽取出相关的实体及实体之间的关系，组成一个结构化的知识库。本文在准备工作阶段通过半自动化模式抽取系统 EARES[20]生成实验知识语料库。本文的工作重点为：进行数据处理以生成正样本语料，运用卷积神经网络进行实体关系分类从而实现关系抽取。整体关系抽取过程如图 1 所示。

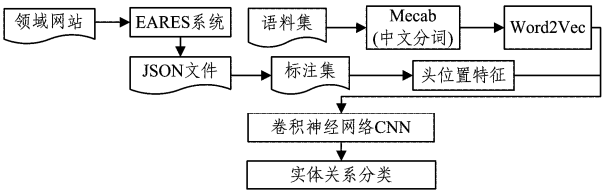

图 1 框架系统关系抽取过程

3.1 数据处理得到正样本语料

针对资源数据大规模和多源异构的特性，一个实用且有效的关系抽取方法应增强正样本训练数据自动获取且减少人工标注数据的耗时以及解决伴随的一致性问题。因此，正样

本训练数据的处理过程主要包含 3 个阶段。

(1)句子级实体识别。在 EARES 系统生成的实验语料中，涵盖部分段落级描述或结构为复杂型的复杂句。因此，为了使大量复杂训练语料转化为易提取的训练语料，需要对数据语料进行预处理。首先根据实验语料中一部分已标注的 JSON 文件数据进行实体抽取，从而构建出本实验所需的医疗术语表；然后对语料中的数据段落或复杂句句型进行切割或拆分，使其转化为能够被计算机识别处理的简单句型。在此过程中可基于已生成的医疗术语表进行词表拓展，对句子中的词语进行分词处理，本实验采用中科院 ICTCLAS 工具来完成该部分的预处理。

(2)标注语句实体关系。弱监督关系抽取样本提取的主要思想在于如何自动生成正样本的训练语料。本文主要基于弱监督假设：根据两个实体在知识库中所具有的关系，推断包含两个实体的语句在知识库中也具有该实体对关系。因此在生成的训练语料中必须包含对相关实体的关系标注。通过系统进行的预处理和词表拓展分词过程，能够比较清晰地得出某一语句中所表述实体关系中的第一实体归属，因此此后在其他实体关系识别中，先以第一个实体为主体，再依据该实体属性寻找该语句中所匹配的属性客体(比如针对“相对症状”这一属性，需要在属性值中寻找症状实体)，然后将所寻找到的实体与最初标识的主体配对并标识出两实体间的实体关系。

弱监督回标关系抽取主要基于以下假设：若两个实体在知识库中具有一定的关系，则根据同时包含这两个实体的句子，推断出实体对在知识库中具有的关系。由于语言表达的多样性，弱监督的这种假设太过武断，两个实体出现在同一个句子中并不能表明它们一定具有某种语义关系，这两个实体可能只是属于同一个话题。虽然弱监督关系抽取的样本提取能够自动生成训练语料，但同时也存在携带噪声的问题。图 2 给出了两个回标语句样本，第一个语句表达实体对关系为相关症状，而第二个语句则无法表述这种关系，这种错误的回标样本即为回标噪声，此回标噪声产生的原因在于弱监督假设对生成的样本数据的约束太少。

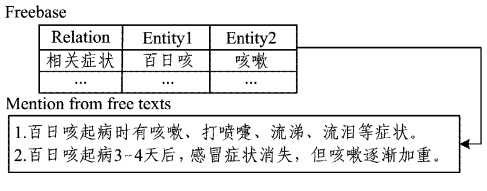

图 2　弱监督关系抽取数据产生过程示例

因此本文在生成数据集时加入如下约束：

1)实体本身作为回标语句中句子的开头，是实体对关系中的第一个实体。

2)根据实体属性，在对应的属性值中识别相匹配的另一个实体，比如针对“相关症状”这一属性，需要在属性值中寻找症状实体。在表 1 中可以通过实体识别脚本找出“咳嗽”、“打喷嚏”和“流涕”等症状实体；而针对“相关检查”这一属性，则需要在属性值中寻找检查实体。

3)在采集文本分句后识别实体时，只将第一个被识别的实体作为实体对关系的第二个实体。并假设：属性值是针对实体属性的描述文本，一般包含实体对关系的另一个实体。

表 1　实体属性关系数据

实体	属性	属性值
百日咳	相关症状	起病时有咳嗽、打喷嚏、流涕、流泪、低热或中度发热等类似感冒症状，3～4 天后症状消失，热退，但咳嗽逐渐加重，尤以夜间为重，此时传染性最强，可持续 7～10 天，若及时治疗，则能有效地控制本病的发展。因缺氧而出现发绀，甚至于抽搐，亦可因窒息而死亡。…
	相关检查	1)血液检查：在卡他期末及痉咳早期白细胞计数高达(20～40)×109/L，最高可达 100×109/L，分类淋巴细胞在 60% 以上，亦有高达 90% 以上者。2)细菌培养：目前认为鼻咽拭培养法优于咳碟法。…
	…	…

最终可以得到类似于如下的标注语句(训练语料句子)：

1:〈e1〉百日咳〈/e1〉因缺氧而出现〈e2〉发绀〈/e2〉，甚至于抽搐，亦可因窒息而死亡。

其中，1 表示标注标签(标签对应关系如表 2 所列)；〈e1〉百日咳〈/e1〉表示标注的实体 1；〈e2〉发绀〈/e2〉表示标注的实体 2。

该语句所要表达的意思为：实体“百日咳”和实体“发钳”之间的关系为标签“1”，即实体对关系：1(百日咳，发绀)或者相关症状(百日咳，发绀)。

表 2　标签对应关系

关系	相关症状	相关疾病	相关检查	并发症	相关治疗
标签	1	2	3	4	5

(3)特征词向量化。自然语言处理的相关任务中，通常需要将语言符号数学化后，才能将自然语言交给机器学习中的算法来处理。词向量是将自然语言中的词表达成数学化的一种方式，它是将一个词表示成一个向量，因此将基于弱监督方法获取得到的训练样本输入到卷积网络之前，需要数字化训练样本。对于图 3 给出的输入示例“〈e1〉百日咳〈/e1〉因缺氧而出现〈e2〉发绀〈/e2〉，甚至于抽搐，亦可因窒息而死亡。”，对句子中的每个词语进行词向量表达，同时记录每个词的头尾位置，并将两个实体的头位置(比如“百日咳”的头位置是 0，“发绀”的头位置是 5)进行位置向量表达。本实验从 Wikipedia 获取到 1.1G 中文预料以及 399M 半自动抽取的医学领域语料，在对所得到的语料进行分词处理后，将得到的分词结果输入到 word2vec 中的 CBOW 模型用以训练维度为 300 的词向量的词嵌入查询表 W。此方法能够通过词嵌入查询表将训练样本中的每个词转换为对应的词向量。此外，将两个实体头位置映射成为实数向量和，得到实体头位置矩阵 H。通过拼接两个实体的词向量表达和实体头位置特征向量矩阵表达，可得到卷积层需要的输入文本向量化表示结果，每个样本示例可以表示为矩阵，如式(1)所示：

$$X=[v_1,v_2,\cdots,v_i,h_{e1},h_{e2}]=[x_1,x_2,\cdots,x_i,\cdots,x_n] \quad (1)$$

其中，$[v_1,v_2,\cdots,v_i]$为训练样本中每输入一个示例句子对应的词向量表达，设为矩阵 V；h_{e1} 和 h_{e2} 为两个实体头位置向量表达，则矩阵 X 的维度为 $d_x+d_k*2(d_k=d_{k1}=d_{k2})$，$d_x$ 为矩阵 X 的维度，d_k 为实体头位置矩阵的维度。

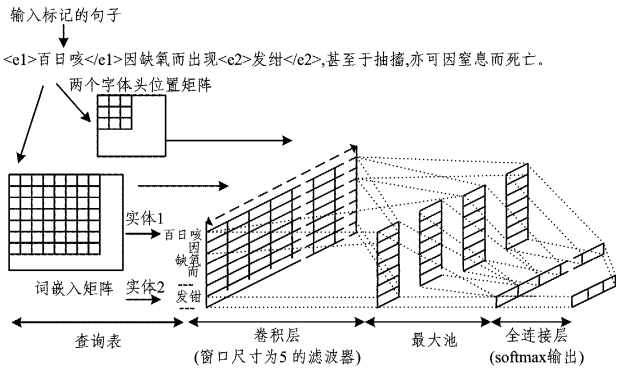

图 3 基于卷积神经网络的关系抽取过程

3.2 基于卷积神经网络的分类器

本文主要是将卷积神经网络用于实体关系的分类器。在将正样本语料进行数字化后，将词向量表达得到的矩阵 X 输入到卷积层，用于自动抽取更高级别的特征。对于池化层，有 N 个输入 maps 就有 N 个输出 maps，只是每个输入 map 的维度都变小了，因为它只保留有效信息来获取最终的显著特征，所以本文使用最大池化。分类操作是卷积神经网络的最后一个操作，首先使用 dropout 进行正则化操作，dropout 在训练网络的过程中可以随机移除一些神经元来保持输入输出层不变，以此来防止过拟合问题；然后通过 softmax 分类器模型对从池化层传递过来的并且经过 dropout 过程的特征 P 进行分类运算。

在 softmax 回归中解决了多分类问题，即针对类标记 y 可以取 k 个不同的值。因此，对于训练集 (p,y)，有 $y\in\{1,2,\cdots,k\}$。例如，在本实验任务中有 5 个不同的类别。对于最终得到的特征 P，可以用假设函数针对每一个类别 j 估算出概率值 $g(y=j:p)$。因此，本文的假设函数将要输出一个 k 维的向量(向量元素的和为 1)来表示这 k 个估计的概率值。具体地说，假设函数 $h_\theta(P)$ 形式化的表示如式(2)所示：

$$h_\theta(p)=\begin{bmatrix} g(y=1:p;\theta) \\ g(y=2:p;\theta) \\ \cdots \\ g(y=k:p;\theta) \end{bmatrix}=\frac{1}{\sum_{j=1}^{k} e^{\theta_j^T p}}\begin{bmatrix} e^{\theta_1^T p} \\ e^{\theta_2^T p} \\ \cdots \\ e^{\theta_k^T p} \end{bmatrix} \tag{2}$$

其中，$\theta_1,\theta_2,\cdots,\theta_k$ 是 softmax 层参数，$\frac{1}{\sum_{j=1}^{k} e^{\theta_j^T p}}$ 是对概率分布进行的归一化操作，使得所有的概率之和为 1。

4 实验与分析

4.1 评价数据及评价指标

本实验使用基于弱监督的样本提取方法，从获取得到的实体属性关系中进一步提取实体关系。本数据集主要是针对疾病实体与其他实体之间的关系语句，总共有 16612 个标注样本，其中 80%作为样本集，20%作为测试集，其中设置了 5 类有向关系，分别为相关症状、相关疾病、相关检查、并发症和相关治疗。其标签对应关系如表 2 所列。

本节使用准确率、召回率和 F1-Score 来评价算法的性能，其中 F1-Score 的计算公式如式(3)所示：

$$F1=2\frac{P\times R}{P+R} \tag{3}$$

其中，$P=\frac{T_p}{T_p+F_p}$ 是准确率，$R=\frac{T_p}{T_p+F_n}$ 是召回率。T_p 表示分类器判定正确并且事实也是正确的信息条数；F_p 表示分类器判定正确但实际错误的信息条数；F_n 表示分类器判定错误但实际正确的信息条数。

4.2 对比实验

实验中的相关参数设置如下：卷积核的窗口大小为 5，卷积核的个数为 150，使用的词向量的维数为 300，实体头位置向量的维数为 10，dropout 的概率为 0.5，模型的迭代次数为 25000。训练得到的模型准确率曲线图如图 4 所示。

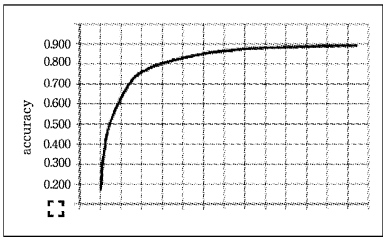

图 4 CNN 模型训练时的准确度曲线图

本实验使用传统的机器学习分类方法(决策树(Decision Tree)、k-近邻(k-Nearest Neighbors)、高斯朴素贝叶斯(Gaussian Naive Bayes)、SVM)与本文方法进行比较，表 3 列出了基于卷积神经网络分类情况的混淆矩阵，表 4－表 7 分别列出了基于决策树、k-近邻、高斯朴素贝叶斯以及 SVM 分类情况的混淆矩阵。从表中可以看出，基于卷积神经网络的实体对关系分类比传统的机器学习方法的误报率低。可以发现第 1 类和第 4 类的分类情况都比较差，原因在于第 1 类是“相关症状”，第 4 类是“并发症”，这两类关系本身的关联性比较强，因此容易出现第 1 类别错分到第 4 类别以及第 4 类别错分到第 1 类别的情况。

表 3 基于卷积神经网络分类情况的混淆矩阵

类别	1	2	3	4	5
1	1047	5	14	40	9
2	3	323	1	7	0
3	25	3	952	3	3
4	217	0	6	351	9
5	10	1	4	3	287

表 4 基于决策树分类情况的混淆矩阵

类别	1	2	3	4	5
1	773	46	20	269	7
2	42	250	12	20	10
3	10	4	962	5	5
4	313	10	4	253	3
5	9	4	12	5	275

表 5 基于 k-近邻分类情况的混淆矩阵

类别	1	2	3	4	5
1	934	9	9	157	6
2	89	207	13	11	14
3	5	0	974	2	5
4	339	5	0	236	3
5	12	0	4	0	289

表 6　基于高斯朴素贝叶斯分类情况的混淆矩阵

类别	1	2	3	4	5
1	630	160	0	323	2
2	4	282	0	48	0
3	8	10	907	52	9
4	188	78	0	317	0
5	0	66	0	0	239

表 7　基于 SVM 分类情况的混淆矩阵

类别	1	2	3	4	5
1	1085	5	1	21	3
2	60	266	1	7	0
3	8	0	972	0	6
4	491	9	0	83	0
5	0	0	3	0	302

本文方法与传统关系抽取方法在准确率、召回率、F1-Score 等性能参数下的比较如表 8 所列。

表 8　本文方法与传统机器学习方法的性能评价

算法	准确率/%	召回率/%	F1-Score
本文方法	91.87	91.58	0.8908
SVM	83.24	81.49	0.7768
决策树	75.33	75.62	0.7546
k-近邻	79.61	79.45	0.7864
高斯朴素贝叶斯	76.19	71.47	0.7258

从表 8 可以看出,本文使用卷积神经网络的方法最终在准确率、召回率以及 F1-Score 性能评价指标上均比传统的机器学习分类方法高出约 10%。传统机器学习的关系抽取方法在医疗训练语料的预处理与特征提取上只能以常规面向中文的方式进行,并没有深入到医疗领域内的实体关系特征,以致在特征提取阶段达不到高准确率,并且回标过程中也包含了大量的回标噪音,这也造成最终准确率与召回率的下降。

结束语　本文实验结果表明,通过优化弱监督的训练正样本和采用卷积网络的分类方式能够有效提高关系抽取的抽取性能。本文主要贡献有:1)通过在弱监督过程添加关系约束及对训练语料进行特征向量矩阵转换,提高了训练语料质量,减少了训练语料的回标噪声对样本的影响;2)通过采用卷积网络的分类方式验证了该方法在弱监督学习方法分类器上的可行性。

参 考 文 献

[1] XU J, ZHANG Z X, WU Z X. Review on Techniques of Entity Relation Extraction[J]. New Technology of Library and Information Service, 2008(8): 1003-3513. (in Chinese)
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(8):1003-3513.

[2] YAO L M, RIEDEL S, MCCALLUM A. Unsupervised relation discovery with sense disambiguation[C]// Proceedings of ACL 2012. 2012: 712-720.

[3] FADER A, SODERLAND S, ETZIONI O. Identifying relations for open information extraction [C]// Proceedings of EMNLP 2011. 2011: 1535-1545.

[4] GRISHMAN R. Information Extraction[M]// Information Retrieval. Springer New York, 2003: 8-15.

[5] KAMBHATLA N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations [C]// ACL 2004 on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, 2004: 22.

[6] ZHOU G D, SU J, ZHANG J, et al. Exploring various knowledge in relation extraction[C]// Proceedings of ACL 2005. 2005: 427-434.

[7] JIANG J, ZHAI C X. A systematic exploration of the feature space for relation extraction[C]// Proceedings of HLT-NAACL 2007. 2007: 113-120.

[8] ZELENKO D, AONE C, RICHARADELLA A. Kernel methods for relation extraction[J]. The Journal of Machine Learning Research, 2003, 3(3): 1083-1106.

[9] ZHANG M, ZHANG J, SU J, et al. Acomposite kernel to extract relation sbetween entities with both flat and structured features[C]// Proceedings of ACL 2006. 2006: 825-832.

[10] ZHOU G D, ZHANG M, JI D H, et al. Tree kernel-based relation extraction with context-sensitive structured parse tree information[C]// Proceedings of EMNLP CoNLL 2007. 2007: 728-736.

[11] CRAVEN M, KUMLIEN J. Constructing biological knowledge bases by extracting information from text sources[C]// Proceedings of AAAI 1999. 1999: 77-86.

[12] RIEDEL S, YAO L M, MCCALLUM A. Modeling relations and their mentions without labeled text[C]// Procedings of ECML PKDD 2010. 2010: 148-163.

[13] SURDEANU M, TIBSHIRANI J, NALLAPATI R, et al. Manning. Multi-instance multi-label learning for relation extraction [C]// Proceedings of EMNLP-CoNLL 2012. 2012: 455-465.

[14] XU W, HOFFMANN R, ZHAO L, et al. Filling knowledge base gaps for distant supervision of relation extraction[C]// Proceedings of ACL 2013. 2013: 665-670.

[15] CHEN Y, GENG G H, JIA H. Density center graph based weakly supervised classification algorithm[J]. Computer Engineering and Applications, 2015, 51(6): 6-10. (in Chinese)
陈燕,耿国华,贾晖. 基于密度中心图的弱监督分类方法[J]. 计算机工程与应用,2015,51(6):6-10.

[16] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12(1): 2493-2537.

[17] YIH W, HE X, Meek C. Semantic Parsing for Single- Relation Question Answering[C]// Proceedings of ACL 2014. 2014: 643-648.

[18] KIM Y. Convolutional neural networks for sentence classification[J]. Eprint Arxiv, 2014: 1408-5882.

[19] ZENG D, LIU K, LAI S, et al. Relation Classification via Convolutional Deep Neural Network[C]// Proceedings of COLING 2014. 2014: 2335-2344.

[20] ZOU Y W, GU J G, FU H D. EARES: Medical Entity and Attribute Extraction System based on Relation Annotation[J]. Wuhan University Journal of Natural Sciences, 2016, 2(21): 145-150.

[21] ZENG D J. Research on Key Technologies of Relational Extraction for Unstructured Text[D]. Beijing: University of Chinese Academy of Sciences, 2015. (in Chinese)
曾道建. 面向非结构化文本的关系抽取关键技术研究[D]. 北京:中国科学院自动化研究所,2015.